

SPRINGER
REFERENCE

Xuemin (Sherman) Shen
Xiaodong Lin
Kuan Zhang
Editors

Encyclopedia of Wireless Networks

Encyclopedia of Wireless Networks

Xuemin (Sherman) Shen • Xiaodong Lin
Kuan Zhang
Editors

Encyclopedia of Wireless Networks

With 455 Figures and 63 Tables

 Springer

Editors

Xuemin (Sherman) Shen
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada

Xiaodong Lin
Institute of Faculty of Business and Information
University of Ontario
Oshawa, ON, Canada

Kuan Zhang
University of Nebraska–Lincoln
Omaha, NE, Canada

ISBN 978-3-319-78261-4 ISBN 978-3-319-78262-1 (eBook)

ISBN 978-3-319-78263-8 (print and electronic bundle)

<https://doi.org/10.1007/978-3-319-78262-1>

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG. The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The evolution of mobile communication systems from the first to the fifth generation (5G) has revolutionized many aspects of society in the past four decades. Wireless networking technologies have become closely integrated with various activities in industry, business, entertainment, and daily life. With improved reliability and increased connection density, wireless networks have been attracting enterprise users, in addition to conventional mobile communication users, by supporting autonomous driving, factory automation, intelligent e-healthcare, smart city, and various other use cases.

Networking services have become highly heterogeneous and dynamic as we continuously integrate wireless networks into a wide variety of applications. Maturing wireless networking technology has attracted ever-increasing interests from multi-disciplines, including communications engineering, electrical and computer engineering, industry and system engineering, social science, economy, and operation management. Tremendous scientific and research efforts are currently aiming at the development of the next generation of heterogeneous, ubiquitous, energy-efficient, sustainable, reliable, and secure wireless networks.

Encyclopedia of Wireless Networks serves to provide comprehensive references to key concepts of wireless networks, including research results of historical significance, areas of current interests, and new directions for future wireless networks. The covered areas range from wireless networking technologies to applications, from the physical layer to the application layer, and from communications to computing in wireless network. This encyclopedia broadly addresses a large cross-section of current development and state-of-the-art technologies from more than 1000 researchers. Entries include descriptions of background, important definition, in-depth illustration, emerging applications, and bibliography. They are easily accessible with user-friendly online search and cross-reference.

The intended audience of this encyclopedia is broad and inclusive. It includes anyone concerned with wireless network technology and its relevant applications. It should be a valuable and authoritative reference for students, teachers, researchers, engineers, and practitioners who need a quick and authoritative reference to a subject in wireless networks and its relevant applications.

We would like sincerely thank the Editorial Board members and individual authors for their significant contribution to Encyclopedia. We also appreciate

Springer's editors and staff, including Susan Lagerstrom-Fife, Juby George, and Rukmani Parameswaran for their continuous support.

Finally, we hope the readers find this encyclopedia interesting and helpful. Any feedback to further improve the encyclopedia will be appreciated.

July 2020

Xuemin (Sherman) Shen
Xiaodong Lin
Kuan Zhang

List of Topics

5G Network

Section Editors: *Rahim Tafazolli and Rose Hu*

5G Wireless
Big Data in 5G
Cellular Unlicensed Spectrum Technology
Future Wireless Network Architecture
and Network Slicing
Hyper Cellular Network: Control-Traffic
Decoupled Radio Access Network
Architecture
Non-orthogonal Multiple Access
Sparse Code Multiple Access
Three-Dimensional Channel Measurement
for the Indoor 5G Multiuser Massive
MIMO System
Wireless Edge Caching
Wireless Fading Channel Model for
5G and Future

Ad hoc Network

Section Editor: *Cheng Li*

Asynchronous Coordination Techniques
Fingerprinting Localization
Hidden and Exposed Terminal Problems
Joint Network Coding and Opportunistic
Forwarding
Location Based Service of Ad Hoc Networks
Multiple Access Techniques
Network Coding-Aware Routing in Wireless
Ad Hoc Networks
Opportunistic Routing (OR)
Routing for Vehicular Ad Hoc Networks

Big Data for Networking

Section Editor: *Song Guo*

Big Data Networks
Big Network Data
Data-Driven Edge Computing
Data-Driven Mobile Social Networks
Data-Driven Network Control
Data-Driven Pattern Prediction
Data-Driven QoE Measurement
Data-Driven Resource Allocation
Data-Driven Security
Data-Driven Service Provisioning
Data-Driven Smart City
Industrial Big Data
Ubiquitous Big Data
Urban Big Data
Wireless Big Data

Cellular Network, 2G/3G Network, 4G/LTE Network

Section Editor: *Hsiao-Hwa Chen*

Analog and Digital Communications
Architectures, Key Techniques, and
Future Trends of Heterogeneous Cellular
Networks
Artificial Noise Schemes Based on MIMO
Technology in Secure Cellular Networks
Base Station Cooperations Under Imperfect
Conditions
Cellular Networks: An Evolution from 1G to 4G
Coordinated Multipoint Transmission
and Reception
Cyclic Prefix-Free OFDM/OFDMA Systems,
Design, and Implementation

Densification of Wireless Networks, from
 Few to Massive
 Deployment Strategies of Cellular Networks
 Efficient Beamforming Design for Cellular
 Networks with Energy-Constrained
 Devices
 Energy Efficiency of Cellular Networks
 Evolution of Vehicular Networks in the
 Transition from 3G to 4G Systems
 Fog Computing for Intelligent Mobile and
 IoT Networks
 Heterogeneous Small Cell Networks
 Key Technologies in 4G/LTE Network
 Machine Learning Paradigms in Wireless
 Network Association
 Mobile Edge Computing: Low Latency
 and High Reliability
 Multiple Access Technique for Cellular
 Wireless Networks
 Random Access Technology in Cellular
 Networks
 Reinforcement Learning-Based Wireless
 Communications Against Jamming
 and Interference
 Relay-Involved Device-to-Device
 Communications in LTE Networks
 Relaying in LTE-Advanced
 Resource Management in Macrocell–Small
 Cell Systems and D2D-Assisted Cellular
 Systems
 Security and Privacy in 4G/LTE Network
 Sidelink Transmissions and Proximity
 Services in LTE-A and Beyond
 Survey on Machine-Learning Techniques
 in LTE Networks
 The Primary Synchronization Sequences
 in LTE

Cognitive Radio Network

Section Editor: *Ning Zhang*

Cognitive Radio Sensor Networks
 Cognitive Radio-Based Non-orthogonal
 Multiple Access
 Cooperative Cognitive Radio Networks
 Database-Based Spectrum Access
 Dynamic Spectrum Access in Multiple
 Primary Networks

Dynamic Spectrum Access Through Cognitive
 Radio Networking
 Energy Harvesting Cognitive Radio Sensor
 Networks
 Game Theory-Assisted Cooperative Spectrum
 Access
 MAC in Cognitive Radio Networks
 Spectrum Resource Allocation in Cognitive
 Radio Networks
 Spectrum Sensing in Cognitive Radio Networks
 Spectrum Sensing with Power Recognition

Cooperative Communications

Section Editor: *Kaoru Ota*

Radio-on-Demand Wireless LAN

Data Center Network

Section Editor: *Lei Lei*

Bandwidth Allocation in Data Center Networks
 Data Center Network in Cloud Computing
 Data Center Network Virtualization
 Data Center Networking (DCN): Infrastructure
 and Operation
 Edge Data Centers in Fog Computing
 Load Balancing in Data Center Networks
 SDN-Enabled Data Center Networks
 Wireless Communication in Data Center

Delay Tolerant and Opportunistic Network

Section Editor: *Yuanguo Bi*

Contact Plan Design for Space Disruption/
 Delay-Tolerant Networks
 Data Dissemination in Delay Tolerant Networks
 with Geographic Information
 Distributed Community Detection in
 Opportunistic Networks
 Mobile Data Offloading via D2D-Assisted
 Communications in Wireless Networks
 Opportunistic Routing in Underwater Sensor
 Networks
 Security in Delay-Tolerant Networks

Equalization, Synchronization and Channel Estimation

Section Editor: *Yingying (Jennifer) Chen*

Equalization Techniques for Single-Carrier Modulations
 Fundamental Properties of Wireless Relays and Their Channel Estimation
 Interference Alignment in Wireless Networks
 Learning-Based Secure Mobile Offloading

Future Network Architecture

Section Editor: *Wei Quan*

Blockchain: Enabling Future Internet with Intrinsic Security
 Collaborative Multipath Transmission
 Computation Offloading in Mobile Edge Computing
 Information-Centric Networking: Basic Principle and Architecture
 Mobile Crowdsensing: Issues and Challenges
 Smart Identifier Network
 Universal Identifier-Based Network

Game Theory in Wireless Network

Section Editor: *Dusit Niyato*

Game Theoretic Modeling of Zero-Rating Internet Provisioning
 Matching Game for Load Distribution in Multi-tier Cellular Networks
 Robust Games for Power Control in Multi-tier Cellular Networks

Interference Characterization and Mitigation

Section Editor: *Lin Cai*

Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space
 Interference Alignment for Power Line Communications
 Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks
 Interference-Aware Distributed Resource Allocation
 Interference-Aware Scheduling in Wireless Networks

Non-orthogonal Multiple Access (NOMA)
 Path Loss and Interference Analysis for a Random Two-Hop Relay Link
 Random Distance Distribution

Internet of Things

Section Editors: *Xiuzhen Cheng and Wei Cheng*

Applications and Architectures of Social Internet of Things
 Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges
 Internet of Things (IoT) Forensic Science
 Internet of Things: Architecture, Key Applications, and Security Impacts
 Maritime Internet of Vessels
 Situation Awareness in Smart Grids

Internet of Things and Its Applications

Section Editor: *Phone Lin*

Anomaly Detection for IoT Systems
 Conversational Chatbot for Disaster Information
 EV Charging Scheduling
 Machine-Type Communication
 Narrowband Internet of Things
 Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies
 Secure Device Pairing
 Social Internet of Things
 Social IoT Crowd-Sourcing on Disaster Reduction

Interworking Heterogeneous Wireless Network

Section Editor: *Ping Wang*

Area Spectral Efficiency of Ultradense Networks
 Cognitive Heterogeneous Networks
 Mobile Edge Caching in HetNets
 Network Selection in Heterogeneous Wireless Networks
 WiFi Offloading in WiFi-Cellular Networks

Medium Access Control

Section Editors: *Hassan Aboubakr Omar* and *Qiang Ye*

Adaptive Medium Access Control for
Internet-of-Things-Enabled
MANETs

Millimeter-wave Communications

Section Editor: *Ming Xiao*

Hybrid Precoding
Lens Antenna Array
Low-Latency Communications with
Millimeter Wave
Massive MIMO BDMA Transmission
Millimeter Wave Beam Training and
Tracking
Millimeter Wave Channel Access
Millimeter Wave Channel Measure
Millimeter Wave Channel Modeling
Millimeter Wave MAC Layer
Millimeter Wave Massive MIMO
Millimeter Wave NOMA
Millimeter Wave Security
Millimeter-Wave (mmWave) Multiple-Input
Multiple-Output (MIMO) Technique
Omnidirectional Transmission for Massive
MIMO
Per-Beam Synchronization for Millimeter-Wave
Massive MIMO
Pilot Reuse for Massive MIMO

MIMO-based Network

Section Editor: *Wei Zhang*

Cell-Free Massive MIMO
Index Modulation
Massive MIMO
Massive MIMO Channel Estimation
MIMO Detection
MIMO Relay
Network MIMO
Satellite MIMO
Space-Time Block Codes
Spatial Modulation

Mobility Management and Models

Section Editors: *Sandra Céspedes* and *Sangheon Park*

Distributed IP Mobility Management
Handoffs, Modeling in IP Networks
Handover in LTE-U/LAA
Mobile IP
Mobility Management in 5G
Mobility Management in NDN
Mobility Management with ID/Locator
Separation
Mobility in Multicast
Network-Layer Mobility Management
Proxy Mobile IPv6

Molecular, Biological and Multi-scale Communications

Section Editor: *Adam Noel*

Applications of Molecular Communication
Systems
Brain-Machine Interfaces
Brownian Motion
Channel Modeling of Microscale Terahertz
Communication
Communications and Networking in
Droplet-Based Microfluidic
Systems
Connectivity via Molecular Signaling
Drug Delivery via Nanomachines
Energy Modeling for Nanoscale Terahertz
Communication
Modeling Approaches for Simulating Molecular
Communications
Modulation in Molecular Signaling
Molecular Bit Detection
Molecular Event Detection
Nanonetworks
Nanoscale Terahertz Communications
Neuronal Communication Channels
Reaction-Diffusion Channels
Receiver Mechanisms for Synthetic
Molecular Communication Systems with
Diffusion
Synchronization of Nanomachines

Network Economics and pricing

Section Editors: *Jianwei Huang and Yuan Luo*

Auction
 Budget Feasible Mechanisms
 Collusion
 Dynamic Pricing
 Efficiency and Pareto Optimality
 Fairness
 Market Entry
 Matching for Cooperative Spectrum Sharing
 Network Effects
 Oligopoly Pricing in Wireless Networks
 Preferences and Utility Functions
 Prospect Theory for Communication Networks
 Sharing Economics

Network Forensics and surveillance, Fault Tolerance and Reliability

Section Editor: *Hongwei Li*

General Theory of Information Security
 Network Privacy Surveillance and Privacy Protection

Network Measurement and Virtualization

Section Editor: *Yusheng Ji*

Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement
 Network Slicing-Enabled Green C-RAN
 Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization
 Resource Scheduling in Virtualized Wireless Networks

Quality of Service, Quality of Experience and Quality of Protection

Section Editors: *Rui L. Aguiar and Yu Cheng*

Long-Term Quality of Protection
 QoE Assessment Models
 QoS in Indoor Localization
 QoS-Aware MAC
 Quality of Service in IEEE 802.11 Networks
 Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes

Resource Allocation and Management

Section Editors: *Junshan Zhang and Nan Cheng*

Cache Placement and Content Delivery Design in Wireless Caching Networks
 Effective Utilization of Unlicensed Spectrum
 Incentive Mechanisms for Mobile Crowdsensing
 Media Access Control for Narrowband Internet of Things: A Survey
 Power Control for Data Transmission in Wireless Networks
 Power Control in Wireless Networks
 Resource Allocation in Industrial IoT
 Resource Allocation in NOMA-Assisted D2D Networks
 Resource Allocation in SDN/NFV-Enabled 5G Networks
 Resource Allocation in UAV-Aided Wireless Networks
 Social Group Utility Maximization

Routing and Multi-cast, Router and Switch Design

Section Editor: *F. Richard Yu*

Energy-Aware and Throughput-Improved Route Selection for Wireless Multi-hop Networks
 Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks
 Handoff Management in Wireless Communication-Based Train Control Systems
 Heterogeneous Networks Through Multi-resources Deployment, Performance Enhancement for
 Integrated System of Networking, Caching, and Computing
 Interference Mitigation in Wireless Communications-Based Train Control (CBTC), Cognitive Control Approach
 Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks
 Mining Task Offloading in Wireless Blockchain Networks
 Mobile Content Distribution Architecture
 Offloading Delay-Tolerable Data Traffic to Connected Vehicle Networks

Secure Routing in Cognitive Radio Vehicular Ad Hoc Networks
 Topology Control in Mobile Ad Hoc Networks with Cooperative Communications
 Trust-Based Routing in Software-Defined Vehicular Ad Hoc Networks

Scaling Laws and Fundamental Limits

Section Editor: *Ning Lu*

Capacity of Wireless Ad Hoc Networks
 Network Information Theory
 Transmission Capacity

Security, Privacy and Trust

Section Editor: *Kui Ren*

Access Control
 Authentication
 Bluetooth Low Energy (BLE) Security and Privacy
 Byzantine Attack and Defense in Wireless Data Fusion
 Encrypted Search for Medical Data
 Flow Table Overflow Attacks
 IoT Security
 Mobile Security and Privacy
 Multimedia Security
 Privacy-Preserving Data Collection
 Private Truth Discovery in Cloud-Empowered Crowdsensing Systems
 Radio-Frequency Fingerprinting-Based Physical Layer Identification
 Vehicle Ad-Hoc Networks: Services, Security, and Privacy
 Verifiable Cloud Computing
 Wireless Key Establishment

Short Range Communications, RFID and NFC

Section Editor: *Zhiguo Shi*

Indoor Visible Light Communication Networks for Camera-Based Mobile Sensing
 Smart Device-to-Device Communication: Communication Establishment and Maintenance
 Wireless Charger Deployment with Communication Constraint

Smart Grid Communications

Section Editor: *Vincent Wong*

Advances in Distribution System Monitoring Architecture and Data Management for Smart Community Information Platform
 Cyber-physical Security in Smart Grid Communications
 Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future
 Intelligence and Trading with Energy Storage and Renewable Generation of Microgrids
 Power Line Communications for Grid Discovery and Diagnostics
 Privacy-Preserving Data Aggregation for Smart Grids
 Smart Grid Communication Based on IEEE 2030 Standard
 Smart Metering, Specific Challenges, and Solution Approaches
 Ultra-reliable and Low-Latency Communications for the Smart Grid

Vehicular Network

Section Editors: *Lian Zhao and Qing Yang*

Delay-Tolerant Network Routing
 Vehicular Ad Hoc Network
 Wireless Access in Vehicular Environment

Video Streaming

Section Editor: *Zhi Liu*

Edge Computing for Real-Time Video Stream Analytics
 Live Video Streaming in Wireless P2P Networks
 QoE Measurement and Assessment of Video Streaming
 Quality of Experience for Wireless Video Streaming
 Scalable Video Streaming
 Video Streaming over Vehicular Networks
 Wireless Video Delivery
 Wireless Video Streaming

Wireless Body Area Network and e-healthcare

Section Editor: *Honggang Wang*

60 GHz in WBAN and mm-Wave Technologies
Data Analytics for Longitudinal Biomedical Data
Internet of Medical Things
Molecular Communication for Wireless Body Area Networks
Sleep Monitoring Using WiFi Signal
Turning Detection in Sandbar Sharks Through Accelerometer Data

Wireless Security

Section Editors: *Haojin Zhu and Jian Shen*

Automotive Security
Blockchain
Differential Privacy
Distributed Storage Security
Fault Tolerance and Reliability of Smart Grids
Group Key Agreement for Wireless Networks
Mixed Network
New Variants of Mirai and Analysis
Pilot Spoofing Attack
Privacy Game
Privacy Preservation for Location-Based Services
Trustworthiness Evaluation
Wi-Fi Protected Access (WPA)
Wireless Group Authentication
Wireless Jamming Attack

Wireless Sensor Network

Section Editors: *Jiming Chen and Ruilong Deng*

Application of Machine Learning in Wireless Sensor Network
Collaborative Beamforming in Wireless Sensor Networks

Cooperative Localization in Wireless Sensor Networks
Data Gathering in Wireless Sensor Networks
Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks
Energy Harvesting Technologies in Wireless Sensor Networks
Extending WSN Lifetime Based on Evolutionary Clustering Algorithm
Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks
Information-Centric Wireless Sensor Networks
Lattice Reduction and Its Applications in Wireless Sensors Network
Localization in 3D Surface Wireless Sensor Networks
Machine Learning in Wireless Sensor Networks for the Internet of Things
Middleware for Wireless Sensor Networks
Parameter Estimation Exploiting WSNs with Arbitrary Topology Geometry
Protocol Stack Architecture for Wireless Sensor Networks
Wireless Sensor Network Security
Wireless Sensor Networks Enabled Urban Traffic Management

WLAN and OFDM

Section Editor: *Xianbin Wang*

Frequency-Domain Equalization in OFDM and Single-Carrier Modulation
Index Modulation for OFDM
PAPR Reduction in OFDM Systems
Principle of OFDM and Multi-carrier Modulations
Time and Frequency Synchronization in OFDM Systems
Time-Domain Synchronous Orthogonal Frequency Division Multiplexing

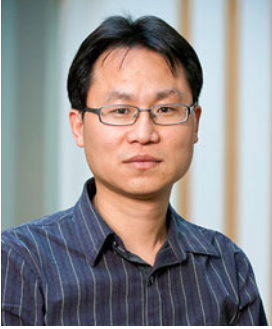
About the Editors



Xuemin (Sherman) Shen (M'97–SM'02–F'09) received his Ph.D. degree in Electrical Engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. Dr. Shen is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular ad hoc and sensor networks. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Dr. Shen received the R.A. Fessenden Award in 2019 from IEEE, Canada, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015, and Education Award in 2017 from the IEEE Communications Society. He has also received the Excellent Graduate Supervision Award in 2006 and Outstanding Performance Award five times from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for the IEEE Globecom'16, the IEEE Infocom'14, the IEEE VTC'10 Fall, the IEEE Globecom'07, the Symposia Chair for the IEEE ICC'10, the Tutorial Chair for the IEEE VTC'11 Spring, and the Chair for the IEEE Communications Society Technical Committee on Wireless

Communications. He was the Editor-in-Chief of the *IEEE Internet of Things Journal* and the Vice President of Publications of the IEEE Communications Society.



Xiaodong Lin is a Ph.D. in Information Engineering from Beijing University of Posts and Telecommunications, China, and the Ph.D. (with Outstanding Achievement in Graduate Studies Award) in Electrical and Computer Engineering from the University of Waterloo, Canada. He is currently an associate professor in the School of Computer Science at the University of Guelph, Canada. Dr. Lin's research interests include computer and network security, computer forensics, privacy protection, applied cryptography, and software security. He is a Fellow of the IEEE.



Kuan Zhang (S'13-M'17) has been an assistant professor at the Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, USA, since September 2017. He received his Ph.D. degree in Electrical and Computer Engineering from the University of Waterloo, Canada, in 2016. Dr. Zhang was also a postdoctoral fellow with the Broadband Communications Research (BBCR) group, Department of Electrical and Computer Engineering, University of Waterloo, Canada, from 2016 to 2017. His research interests include security and privacy for mobile social networks, e-healthcare, cloud/edge computing, and Internet of Things. Dr. Zhang received IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence – Outstanding Ph.D. Thesis – in 2017. He was the recipient of Best Paper Award in IEEE WCNC 2013, Securecomm 2016, and BigDataSE 2019.

Area Editors

Area: 5G Network



Rahim Tafazolli

University of Surrey, UK



Rose Hu

Electrical and Computer Engineering
Department
Utah State University
Logan, UT, USA

Area: Ad hoc Network



Cheng Li

Head of the Department of Electrical and
Computer Engineering
Faculty, Engineering and Applied Science
Memorial University
Canada

Area: Big Data for Networking**Song Guo**

Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Area: Cellular Network, 2G/3G Network, 4G/LTE Network**Hsiao-Hwa Chen**

Department of Engineering Science
National Cheng Kung University
Taiwan

Area: Cognitive Radio Network**Ning Zhang**

Department of Computing Sciences
Texas A&M University at Corpus Christi
TX, USA

Area: Cooperative Communications**Kaoru Ota**

Department of Sciences and Informatics
Muroran Institute of Technology
Hokkaido, Japan

Area: Data Center Network**Lei Lei**

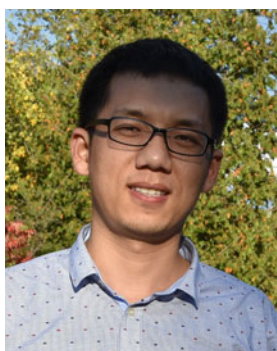
College of Engineering & Physical Sciences
University of Guelph
Ontario, Canada

Area: Delay Tolerant and Opportunistic Network**Yuanguo Bi**

School of Computer Science and Engineering
Northeastern University
Shenyang, China

Area: Equalization, Synchronization and Channel Estimation**Yingying (Jennifer) Chen**

Rutgers University
New Brunswick, NJ, USA

Area: Future Network Architecture**Wei Quan**

National Engineering Lab for Next Generation
Internet Technologies
School of Electronic and Information
Engineering
Beijing Jiaotong University
Beijing, China

Area: Game Theory in Wireless Network**Dusit Niyato**

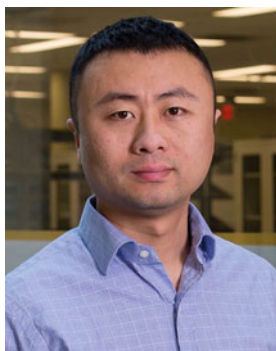
School of Computer Science and Engineering
Nanyang Technological University
Singapore

Area: Interference Characterization and Mitigation**Lin Cai**

Dept. of Electrical & Computer Engineering
University of Victoria
Victoria, BC, Canada

Area: Internet of Things**Xiuzhen Cheng**

Department of Computer Science
The George Washington University
Washington DC, USA

**Wei Cheng**

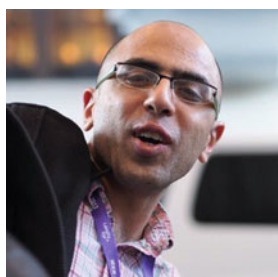
School of Engineering and Technology
University of Washington
Tacoma, WA, USA

Area: Internet of Things and its Applications**Phone Lin**

Computer Science & Information Engineering
National Taiwan University
Taiwan, R.O.C.

Area: Interworking Heterogeneous Wireless Network**Ping Wang**

Department of Electrical Engineering &
Computer Science, Lassonde School of
Engineering
York University
Toronto, Canada

Area: Medium Access Control**Hassan Aoubakr Omar**

Huawei Technologies Inc.
Kanata, ON, Canada

**Qiang Ye**

Department of Electrical and Computer
Engineering and Technology
Minnesota State University
Mankato, MN, USA

Area: Millimeter-wave Communications**Ming Xiao**

Division of Information Science and
Engineering, School of Electrical
Engineering and Computer Science
Royal Institute of Technology
KTH, Sweden

Area: MIMO-based Network**Wei Zhang**

School of Electrical Engineering and
Telecommunications
The University of New South Wales
Sydney, Australia

Area: Mobility Management and Models**Sandra Céspedes**

Department of Electrical Engineering
Universidad de Chile
Santiago, Chile

**Sangheon Pack**

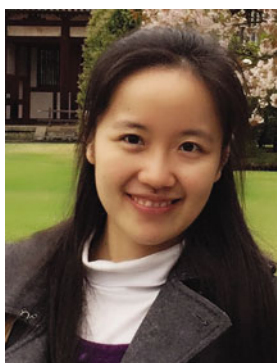
School of Electrical Engineering
Korea University
Seoul, Korea

Area: Molecular, Biological and Multi-scale Communications

Adam Noel
School of Engineering
University of Warwick
Coventry, UK

Area: Network Economics and pricing

Jianwei Huang
School of Science and Engineering
The Chinese University of Hong Kong
Shenzhen



Yuan Luo
Department of Computing
Imperial College
London, UK

Area: Network Forensics and surveillance, Fault Tolerance and Reliability**Hongwei Li**

University of Electronic Science and Technology
of China
Chengdu, China

Area: Network Measurement and Virtualization**Yusheng Ji**

Information Systems Architecture Science
Research Division
National Institute of Informatics
Tokyo, Japan

Area: Quality of Service, Quality of Experience and Quality of Protection**Rui L. Aguiar**

Instituto de Telecomunicações
Universidade de Aveiro
Portugal

**Yu Cheng**

Department of Electrical and Computer
Engineering
Illinois Institute of Technology
Chicago, IL, USA

Area: Resource Allocation and Management

Junshan Zhang
School of ECEE
Arizona State University
Tempe, AZ, USA



Nan Cheng
State Key Lab. of ISN, and School of
Telecommunications Engineering
Xidian University
Xian, China

Area: Routing and Multi-cast, Router and Switch Design

F. Richard Yu
Carleton University
Canada

Area: Scaling Laws and Fundamental Limits

Ning Lu
Department of Electrical and Computer
Engineering
Queen's University
Kingston, Ontario, Canada

Area: Security, Privacy and Trust**Kui Ren**

College of Computer Science and Technology
Zhejiang University
Hangzhou, China

Area: Short Range Communications, RFID and NFC**Zhiguo Shi**

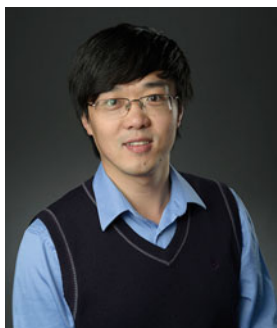
College of Information Science and Electronic
Engineering
Zhejiang University
Hangzhou, China

Area: Smart Grid Communications**Vincent Wong**

Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC, Canada

Area: Vehicular Network**Lian Zhao**

Department of Electrical, Computer &
Biomedical Engineering
Ryerson University
Toronto, Canada

**Qing Yang**

Department of Computer Science and
Engineering Discovery Park
University of North Texas
Denton, TX, USA

Area: Video Streaming**Zhi Liu**

Department of Mathematical and Systems
Engineering
Shizuoka University
Johoku Hamamatsu, Japan

Area: Wireless Body Area Network and e-healthcare**Honggang Wang**

Electrical and Computer Engineering
Department
University of Massachusetts
Dartmouth, MA, USA

Area: Wireless Security**Haojin Zhu**

Department of Computer Science & Engineering
Shanghai Jiao Tong University
China

**Jian Shen**

College of Computer and Software
Nanjing University of Information Science and
Technology
Nanjing, China

Area: Wireless Sensor Network**Jiming Chen**

College of Control science and Engineering
Zhejiang University
Hangzhou, China

**Ruilong Deng**

College of Control Science and Engineering
Zhejiang University
Hangzhou, China

Area: WLAN and OFDM**Xianbin Wang**

Department of Electrical and Computer
Engineering
Western University
London, Ontario, Canada

List of Contributors

Raed Abdullah Hydro Ottawa, Ottawa, ON, Canada

Saeid Aghaeinezhadfirouzja Department of Electronics Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China

Arman Ahmadzadeh Institute for Digital Communications, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany

Arslan Ahmed Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Özgür Barış Akan University of Cambridge, Cambridge, UK

Ian F. Akyildiz Broadband Wireless Networking (BWN) Laboratory, Georgia Institute of Technology, Atlanta, GA, USA

Ahmad Alagil Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

Eyhab Al-Masri School of Engineering and Technology, University of Washington Tacoma, Tacoma, WA, USA

M. D. Nashid Anjum Department of Electrical Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA

Catalina Aranzazu-Suescun Florida Atlantic University, Boca Raton, FL, USA

Omid Ardakanian Department of Computing Science, University of Alberta, Edmonton, AB, Canada

Dizem Arifler Middle East Technical University, Northern Cyprus Campus, Turkey

Dogu Arifler Eastern Mediterranean University, Northern Cyprus, Turkey

Masoud Asghari Department of Computer Engineering, Urmia University, Urmia, Iran

Md. Abul Kalam Azad Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

Lin Bai School of Electronic and Information Engineering, Beijing Laboratory for General Aviation Technology (Beihang University), Beijing, China

- Indika A. M. Balapuwaduge** University of Agder, Kristiansand, Norway
- Sasitharan Balasubramaniam** Telecommunication Software and Systems Group, Waterford Institute of Technology, Waterford, Ireland
- Xuecai Bao** Nanchang Institute of Technology, Nanchang, China
- Carlos J. Bernardos** Universidad Carlos III de Madrid, Madrid, Spain
- Randall A. Berry** Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA
- Emil Björnson** Department of Electrical Engineering (ISY), Linköping University, Linköping, Sweden
- Zied Boudia** Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
- Robert A. Bridges** Cyber and Applied Data Analytics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
- Stefan Bruendl** University of Massachusetts Dartmouth, North Dartmouth, MA, USA
- Sherif Adeshina Busari** Instituto de Telecomunicações, Aveiro, Portugal
- Lin Cai** The Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada
- E&CE Department, Illinois Institute of Technology, Chicago, IL, USA
- Zhipeng Cai** Department of Computer Science, Georgia State University, Atlanta, GA, USA
- Bin Cao** EIE School, Harbin Institute of Technology, Shenzhen, China
- Jiahao Cao** Tsinghua University, Beijing, China
- Xianghui Cao** Southeast University, Nanjing, China
- Yangjie Cao** School of Software and Applications Technologies, Zhengzhou University, Zhengzhou, China
- Yuanlong Cao** Jiangxi Normal University, Nan Chang, China
- Yue Cao** School of Computing and Communications, Lancaster University, Lancaster, UK
- Mihaela Cardei** Florida Atlantic University, Boca Raton, FL, USA
- Jason M. Carter** Cyber and Applied Data Analytics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
- Antonio Caruso** Department of Mathematics and Physics 'Ennio De Giorgi, University of Salento, Lecce, Italy
- Sandra Céspedes** Department of Electrical Engineering, Universidad de Chile, Santiago, Chile

Tzu-Yin Chang National Science and Technology Center for Disaster Reduction, Taipei, Taiwan, ROC

Han-Chieh Chao Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan

Department of Computer Science and Information Engineering, National Ilan University, Yilan, Taiwan

Chao-Yu Chen Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan

Chen Chen Xidian University, Xi'an, Shaanxi, China

Guihai Chen Nanjing University, Nanjing, Jiangsu, China

Jian Chen School of Telecommunications Engineering, Xidian University, Shaanxi, China

Jiasi Chen University of California, Riverside, CA, USA

Shuangwu Chen University of Science and Technology of China, Hefei, China

Shuyi Chen Harbin Institute of Technology, Harbin, China

Si Chen West Chester University, West Chester, PA, USA

Siguang Chen Nanjing University of Posts and Telecommunications, Nanjing, China

Qian Chen Electrical and Computer Engineering, University of Texas, San Antonio, TX, USA

Wen Chen Shanghai Institute of Advanced Communications and Data Sciences, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

Xianfu Chen VTT Technical Research Centre of Finland Ltd, Oulu, Finland

Xiangyu Chen The University of Hong Kong, Hong Kong, China

Xu Chen School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

Yanjiao Chen School of Computer Science, Wuhan University, Wuhan, People's Republic of China

Yi Chen University of Michigan-Shanghai Jiao Tong University Joint Institute (UM-SJTU JI), Shanghai Jiao Tong University, Shanghai, China

Yifan Chen The University of Waikato, Hamilton, New Zealand

Yuanzhu Chen Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Xiang Cheng State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing, China

Man Hon Cheung Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China

Chia-Feng Chiang Information Division, National Science and Technology Center for Disaster Reduction, New Taipei City, Taiwan

Yi-Han Chiang Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo, Japan

Wei-Che Chien Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan, Republic of China

Francesco Chiti Department of Information Engineering, University of Florence, Florence, Italy

Hsin-Hung Cho Department of Computer Science and Information Engineering, National Ilan University, Yilan, Taiwan

Jinho Choi School of Information Technology, Deakin University, Burwood, Australia

Xiaoli Chu Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, UK

Yao-Liang Chung Department of Communications, Navigation and Control Engineering, National Taiwan Ocean University, Keelung, Taiwan

Costas Courcoubetis Engineering Systems and Design, Singapore University of Technology and Design, Singapore, Singapore

Daniel P. Crear Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, VA, USA

Tom Cruz Intel Corporation, Santa Clara, CA, USA

Qimei Cui Beijing University of Posts and Telecommunications, Beijing, China

Haipeng Dai Nanjing University, Nanjing, Jiangsu, China

Linglong Dai Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China

Yanpeng Dai Xidian University, Xi'an, China

Ngoc Dung Dao Huawei Technologies Canada Co., Ltd., Ottawa, Canada

Wouter Deconinck Physics, College of William and Mary, Williamsburg, VA, USA

Flavia C. Delicato DCC-IM/PESC/COPPE – Federal University of Rio de Janeiro, Cidade Universitária, Rio de Janeiro, Brazil

Ilker Demirkol Department of Telematics Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain

Der-Jiunn Deng Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua City, Taiwan

Ruilong Deng Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

Yansha Deng Department of Informatics, King's College London, London, UK

Rajib Dey Department of EECS, University of Central Florida, Orlando, FL, USA

Ergin Dinc University of Cambridge, Cambridge, UK

Guoru Ding College of Communications Engineering, Army Engineering University of PLA, Nanjing, China

Zhiguo Ding School of Electrical and Electronic Engineering, University of Manchester, Manchester, UK

Rui Dinis Department of Electrical Engineering, Instituto de Telecomunicações and FCT-UNL, Caparica, Portugal

Lingjie Duan Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore

Melike Erol-Kantarci School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

Hua Fang University of Massachusetts (UMass) Dartmouth, North Dartmouth, MA, USA

UMass Medical School, Worcester, MA, USA

Xuming Fang Southwest Jiaotong University, Chengdu, China

Hamid Farmanbar Huawei Technologies Canada Co., Ltd., Ottawa, Canada

Javad Fattahi University of Ottawa, Ottawa, ON, Canada

L. Felicetti Department of Engineering, University of Perugia, Perugia, Italy

M. Femminella Department of Engineering, University of Perugia, Perugia, Italy

Ransheng Feng School of Computer and Information, Hefei University of Technology, Hefei, Anhui, China

Carlo Fischione KTH Royal Institute of Technology, Stockholm, Sweden

Chuan Heng Foh 5G Innovation Centre, Institute for Communication Systems (ICS), University of Surrey, Guildford, UK

Jun Fu Future Forum, Beijing, China

Xinwen Fu University of Massachusetts Lowell, Lowell, MA, USA

Department of EECS, University of Central Florida, Orlando, FL, USA

Takuya Fujihashi Graduate School of Science and Engineering, Ehime University, Ehime, Japan

Lin Gao School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China

Longxiang Gao Deakin University, Burwood, VIC, Australia

Xinyu Gao Department of Electronic Engineering, Tsinghua University, Beijing, China

Xiqi Gao National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Xiaohu Ge School of Electronic Information and Communications, Director of China International Joint Research Center of Green Communications and Networking, Huazhong University of Science and Technology, Wuhan, China

Hossein S. Ghadikolaei Department of Networks and Systems Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Jie Gong School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

Shimin Gong Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Xiaowen Gong Auburn University, Auburn, AL, USA

Zijun Gong Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

Maxime Grau 5G Innovation Centre, Institute for Communication Systems (ICS), University of Surrey, Guildford, UK

Lin Gu SCTS, CGCL, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

Yu Gu Beijing University of Posts and Telecommunications, Beijing, China

Quansheng Guan School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

João Guerreiro Instituto de Telecomunicações, Lisbon, Portugal
Universidade Autónoma de Lisboa, Lisbon, Portugal

Sri Gundavelli Cisco Systems, San José, CA, USA

Bingjiang Guo College of Intelligence and Computing, Tianjin University, Tianjin, China

Dongchao Guo Beijing Information Science and Technology University, Beijing, P. R. China

Song Guo Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Weisi Guo School of Engineering, University of Warwick, Coventry, UK

Zongming Guo WangXuan Institute of Computer Technology, Peking University, Beijing, China

Biao Han National University of Defense Technology, Changsha, China

Chong Han University of Michigan-Shanghai Jiao Tong University Joint Institute (UM-SJTU JI), Shanghai Jiao Tong University, Shanghai, China

Shuai Han School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China

Ying Han School of Software and Applications Technologies, Zhengzhou University, Zhengzhou, China

Gerhard P. Hancke Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Kun Hao Department of Engineering and Applied Science, Memorial University of Newfoundland, St John's, NL, Canada

Tianjin Chengjian University, Tianjin, China

Takahiro Hara Osaka University, Suita, Japan

Werner Haselmayr Johannes Kepler University Linz, Linz, Austria

Ammar Hawbani University of Science and Technology of China, Hefei, China

Jianping He The Department of Automation, Shanghai Jiao Tong University, and the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China

Shibo He The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

Xiaoming He College of IoT, Nanjing University of Posts and Telecommunications, Nanjing, China

Ying He Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Geoffrey Holmes The University of Waikato, Hamilton, New Zealand

Jia Hou Department of Communications Engineering, Soochow University, Suzhou, Jiangsu, China

Lu Hou Intelligent Computing and Communications Lab (IC² Lab), Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Hsu-Chun Hsiao Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Ching-En Hsieh National Science and Technology Center for Disaster Reduction, Taipei, Taiwan, ROC

Han Hu School of Information and Electronics, Beijing Institute of Technology, Beijing, China

Jiankun Hu School of Engineering and Information Technology, ADFA, Canberra, ACT, Australia

Miao Hu School of Data and Computer Science and Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China

Zhiwen Hu School of Electronics Engineering and Computer Science, Peking University, Beijing, China

Chih-Wei Huang Department of Communication Engineering, National Central University, Taoyuan, Taiwan

Huawei Huang School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

Jie Huang Shandong University, Qingdao, China

Liang Huang College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

Minghe Huang Jiangxi Normal University, Nanchang, China

Shih-Yun Huang Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan

Shuhan Huang College of Intelligence and Computing, Tianjin University, Tianjin, China

Wei Huang Southeast University, Nanjing, China

Xin-Lin Huang Tongji University, Shanghai, China

Yongming Huang School of Information Science and Engineering, Southeast University, Nanjing, China

Zhichuang Huang School of Data and Computer Science and Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China

Kazi Mohammed Saidul Huq Instituto de Telecomunicações, Aveiro, Portugal

Mohamed Ibne Kahla Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Vahid Jamali Institute for Digital Communications, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany

Abdallah Jarwan Carleton University, Ottawa, ON, Canada

Seil Jeon Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea

Hong Ji School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, China

Yusheng Ji Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo, Japan

Juncheng Jia Soochow University, Suzhou, China

Chunxiao Jiang Department of Electronic Engineering, Tsinghua Space Center, Tsinghua University, Beijing, People's Republic of China

Fan Jiang Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

Hai Jiang Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

Ruobing Jiang Ocean University of China, Qingdao, China

Tao Jiang Huazhong University of Science and Technology, Wuhan, China

Xiaofeng Jiang University of Science and Technology of China, Hefei, China

Miao Jin The Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, LA, USA

Shi Jin National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Yier Jin University of Florida, Gainesville, FL, USA

Carlee Joe-Wong Electrical and Computer Engineering, Carnegie Mellon University, Silicon Valley Campus, Moffett Field, CA, USA

Jeongho Jeon Intel Corporation, Santa Clara, CA, USA

Josep M. Jornet University at Buffalo, The State University of New York, Buffalo, NY, USA

Josep Miquel Jornet Ultra-broadband Nano Communication and Networking Laboratory, University at Buffalo, The State University of New York, Buffalo, NY, USA

Somayeh Kafaie Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Canada

Ved P. Kafle Network System Research Institute, National Institute of Information and Communications Technology, Koganei, Tokyo, Japan

Nevrus Kaja University of Michigan, Dearborn, MI, USA

Yichi Kawamoto Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Tooba Khan Koç University, Istanbul, Turkey

Dongmyoung Kim AIX Center, SK Telecom Jung-gu, Seoul, Korea

Hongseok Kim Department of Electronic Engineering, Sogang University, Seoul, South Korea

Chow-In Ko Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan, Republic of China

Haneul Ko Korea University, Seoul, South Korea

Han-Bae Kong School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

Angeliki Kritikakou Univ Rennes, INRIA, CNRS, IRISA, Rennes, Cedex, France

Yonghong Kuo School of Telecommunications Engineering, Xidian University, Shaanxi, China

M. Şükrü Kuran Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey

Chin-Feng Lai Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan, Republic of China

Lutz Lampe Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada

Tian Lan Electrical and Computer Engineering, George Washington University, Washington, DC, USA

Erik G. Larsson Department of Electrical Engineering (ISY), Linköping University, Linköping, Sweden

Lehlogonolo P. I. Ledwaba Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Chia-Peng Lee Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

Jaewook Lee Korea University, Seoul, South Korea

Jong-Hyouk Lee Department of Software, Sangmyung University, Cheonan, Republic of Korea

Mark Leeson University of Warwick, Coventry, UK

Gang Lei Jiangxi Normal University, Nan Chang, China

Xianfu Lei Southwest Jiaotong University, Chengdu, China

Ka-Cheong Leung The University of Hong Kong, Hong Kong, China

Victor C. M. Leung The University of British Columbia, Vancouver, BC, Canada

Bo Li Department of Communication Engineering, Harbin Institute of Technology, Weihai, China

Cheng Li Faculty of Engineering and Applied Science, Memorial University, St. John's, NL, Canada

Fan Li Beijing Institute of Technology, Beijing, China

Fei Li BUPT, Beijing, China

Fuliang Li Northeastern University, Shenyang, China

Frank Y. Li University of Agder, Kristiansand, Norway

Hao Li Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada

He Li Department of Information and Electronic Engineering, Muroran Institute of Technology Muroran, Japan

Jie Li School of Computer and Information, Hefei University of Technology, Hefei, Anhui, China

Junling Li University of Waterloo, Waterloo, ON, Canada

Longpeng Li College of Intelligence and Computing, Tianjin University, Tianjin, China

Meng Li Beijing Institute of Technology, Beijing, China

Pan Li Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

Peng Li School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu, Fukushima, Japan

Qi Li Tsinghua University, Beijing, China

Qiyue Li School of Computer and Information, Hefei University of Technology, Hefei, Anhui, China

Tian Li The 54th Research Institute of CETC, Shijiazhuang Hebei, China

Ting-Mei Li Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan

You Li National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China

Wei Li National University of Defense Technology, Changsha, China

Xi Li School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, China

Xiuhua Li Chongqing University, Chongqing, P. R. China

The University of British Columbia, Vancouver, BC, Canada

Xu Li Huawei Technologies Canada Co., Ltd., Ottawa, Canada

Zan Li Xidian University, Xi'an, China

Ziyang Li Shanghai Institute of Advanced Communications and Data Sciences, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

Chengchao Liang Department of System and Computer Engineering, Carleton University, Ottawa, ON, Canada

Fan Liang Department of Computer Engineering, Towson University, Towson, MD, USA

Hao Liang Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

Wei Liang Northwestern Polytechnical University, Xi'an, China

Shao-Yu Lien National Chung Cheng University, Chia-Yi, Taiwan

Bin Lin School of Information Technology, Dalian Maritime University, Dalian, China

Feng Lin Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO, USA

Hai Lin Department of Electrical and Electronics Systems, Osaka Prefecture University, Sakai, Osaka, Japan

Hong-Chi Lin School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, Hei Longjiang, China

Lin Lin Tongji University, Shanghai, China

Phone Lin Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

Xin-Xue Lin Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Yi-Bing Lin Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

Zhen Ling School of Computer Science and Engineering, Southeast University, Nanjing, China

Alex X. Liu Nanjing University, Nanjing, Jiangsu, China

Chih-Hao Liu Information Division, National Science and Technology Center for Disaster Reduction, New Taipei City, Taiwan

Dengzhi Liu Nanjing University of Information Science and Technology, Nanjing, China

Gongliang Liu Department of Communication Engineering, Harbin Institute of Technology, Weihai, China

Kuang-Hao Liu Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

Mengting Liu School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Mengdi Liu College of Intelligence and Computing, Tianjin University, Tianjin, China

Qiuming Liu Department of Software Engineering, Jiangxi University of Science and Technology, Nanchang, Jiangxi, China

Qinghua Liu Jiangxi Normal University, Nan Chang, China

Shiwen Liu Intelligent Computing and Communication Lab, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Xin Liu Dalian University of Technology, Dalian, China

Xiqing Liu Department of Engineering Science, National Cheng Kung University (NCKU), Tainan City, Taiwan

Yang Liu State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Yao Liu Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

Yiming Liu School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, China

Yi-Liang Liu Department of Electronics Information Engineering, Harbin Institute of Technology, Harbin, China

Zhi Liu Department of Mathematical and Systems Engineering, Shizuoka University, Hamamatsu, Japan

Keping Long Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services, University of Science and Technology Beijing, Beijing, China

Wenjing Lou Virginia Tech, Blacksburg, VA, USA

Yi Lou College of Underwater Acoustic Engineering, Harbin Engineering University, Harbin, China

Tom H. Luan Xidian University, Xi'an, China

Ling Lyu School of Information Science and Technology, Dalian Maritime University, Dalian, China

Di Ma University of Michigan, Dearborn, MI, USA

Lichuan Ma Xidian University, Xi'an, China

Ni Ma Huawei Technologies Co., Ltd., Shenzhen, China

Ruofei Ma Department of Communication Engineering, Harbin Institute of Technology, Weihai, China

Behrouz Maham Electrical and Electronic Engineering Department, Nazarbayev University, Astana, Kazakhstan

Mohammad Upal Mahfuz Richard J. Resch School of Engineering, University of Wisconsin-Green Bay, Green Bay, WI, USA

Guoqiang Mao University of Technology Sydney, Sydney, NSW, Australia

Shiwen Mao Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA

Jon W. Mark E&CE Department, University of Waterloo Waterloo, ON, Canada

Weixiao Meng Harbin Institute of Technology, Harbin, China

Xin Meng Huawei Technologies Co., Ltd., Shanghai, China

Lei Mo Univ Rennes, INRIA, CNRS, IRISA, Rennes, Cedex, France

Muhammad Baqer Mollah Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

Giacomo Morabito University of Catania, Catania, Italy

Reza Mosayebi Sharif University of Technology, Tehran, Iran

Nour Moustafa School of Engineering and Information Technology, ADFA, Canberra, ACT, Australia

Shahid Mumtaz Instituto de Telecomunicações, Aveiro, Portugal

Ross Murch The Hong Kong University of Science and Technology, Hong Kong, China

Zhenyu Na School of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China

Tadashi Nakano Osaka University, Suita, Japan

Nima Namvar Department of Electrical and Computer Engineering, North Carolina A&T State University, Greensboro, NC, USA

Ahmad Nasser University of Michigan, Dearborn, MI, USA

Hien Quoc Ngo School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK

Kejie Ni College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

Zhaolong Ning Dalian University of Technology, Dalian, China

Hiroaki Nishi Department of System Design Engineering, Keio University, Yokohama, Japan

Zhisheng Niu Tsinghua University, Beijing, China

Dusit Niyato School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore, Singapore

Yutaka Okaie Osaka University, Suita, Japan

Sangheon Pack Korea University, Seoul, South Korea

Arindam Pal TCS Research and Innovation, Kolkata, India

Jianping Pan University of Victoria, Victoria, BC, Canada

Ai-Chun Pang Department of Computer Science and Information Engineering, Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan

Federico Passerini Institute of Networked and Embedded Systems, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Qingqi Pei Xidian University, Xi'an, China

Changgen Peng Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang, China

Dan Peng Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Kai Peng College of Engineering, Huaqiao University, Quanzhou, Fujian, P. R. China

Mugen Peng Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing, China

Xuesen Peng Nanjing University, Nanjing, China

Charles Perkins Huawei Technologies, Santa Clara, CA, USA

Massimiliano Pierobon University of Nebraska-Lincoln, Lincoln, NE, USA

Paulo F. Pires DCC-IM/PESC/COPPE – Federal University of Rio de Janeiro, Cidade Universitária, Rio de Janeiro, Brazil

Benjamin D. Powell Computer Science, College of William and Mary, Williamsburg, VA, USA

Jiaju Qi BUPT, Beijing, China

Liping Qian College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

Zhuzhong Qian State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Zhan Qin Texas University at San Antonio, San Antonio, TX, USA

Wei Quan Beijing Jiaotong University, Beijing, China

Atta ul Quddus 5G Innovation Centre, Institute for Communication Systems (ICS), University of Surrey Guildford, UK

Md. Jahidur Rahman Qualcomm Technologies, Inc., San Diego, CA, USA

Hamideh Ramezani University of Cambridge, Cambridge, UK

G. Realì Department of Engineering, University of Perugia, Perugia, Italy

Ju Ren School of Information Science and Engineering, Central South University, Changsha, China

Zhiyuan Ren Xidian University, Xi'an, Shaanxi, China

Michael Rice Department of Electrical and Computer Engineering, Brigham Young University, Provo, UT, USA

Jonathan Rodriguez Instituto de Telecomunicações, Aveiro, Portugal
University of South Wales, Pontypridd, UK

Yue Rong School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA, Australia

Sushmita Ruj Indian Statistical Institute, Kolkata, India

Ayman Sabbah Spectrum and Telecommunication Sector (STS) - Innovation, Science, and Economic Development (ISED)/Government of Canada, Ottawa, ON, Canada

Robert Schober Institute for Digital Communications, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany

Henry Schriemer University of Ottawa, Ottawa, ON, Canada

Nimal Gamini Senarath Huawei Technologies Canada Co., Ltd., Ottawa, Canada

Olivier Sentieys Univ Rennes, INRIA, CNRS, IRISA, Rennes, Cedex, France

Hamed Shah-Mansouri Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada

Hangguan Shan Zhejiang University, Hangzhou, China

Adnan Shaout University of Michigan, Dearborn, MI, USA

Neda Sharifi The University of Waikato, Hamilton, New Zealand

Bobby Sharma School of Technology, Assam Don Bosco University, Guwahati, India

Hang Shen Department of Computer Science and Technology, Nanjing Tech University, Nanjing, Jiangsu Province, China

Xuemin (Sherman) Shen Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

Alaa Sheta Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

Jia Shi Xidian University, Xi'an, China

Junling Shi College of Computer Science and Engineering, Northeastern University, Shenyang, China

Tuo Shi School of Computer Science and Tech, Harbin Institute of Technology, Harbin, China

- Weisen Shi** University of Waterloo, Waterloo, ON, Canada
- Xiufang Shi** Zhejiang University of Technology, Hangzhou, China
- Zhiguo Shi** Zhejiang University, Hangzhou, China
- Yuan-Yao Shih** Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan
- Pengbo Si** Faculty of Information Technology, Beijing University of Technology, Beijing, China
- Basma Solaiman** New and Renewable Energy Authority, Cairo, Egypt
- Jian Song** Tsinghua University, Beijing, China
- Lingyang Song** School of Electronics Engineering and Computer Science, Peking University, Beijing, China
- Mei Song** School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China
- Wei Song** University of New Brunswick, Fredericton, NB, Canada
- Michal K. Stachowiak** Department of Pathology and Anatomical Sciences, University at Buffalo, The State University of New York, Buffalo, NY, USA
- Ruoyu Su** School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China
- Weifeng Su** Department of Electrical Engineering, State University of New York (SUNY) at Buffalo, Buffalo, NY, USA
- Wei Su** Beijing Jiaotong University, Beijing, China
- Wen-Ray Su** National Science and Technology Center for Disaster Reduction, Taipei, Taiwan, ROC
- Shinya Sugiura** Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Koganei, Japan
- Dongun Suh** Korea University, Seoul, Korea
- Kalika Suksomboon** Future Communication System Laboratory, KDDI Research Inc, Fujimino, Saitama, Japan
- Chen Sun** National Mobile Communications Research Laboratory, Southeast University, Nanjing, China
- Haijian Sun** Utah State University, Logan, UT, USA
- Kun Sun** George Mason University, Fairfax, VA, USA
- Ruijin Sun** Pengcheng Laboratory, Shenzhen, China
- Yaohua Sun** Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing, China
- Yuyi Sun** The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

Zhi Sun Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA

Sanaa Taha Information Technology Department, Faculty of Computers and Information, Cairo university, Cairo, Egypt

Suhua Tang Department of Computer and Network Engineering, The University of Electro-Communications, Chofu, Japan

Yanqun Tang School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China

Yujie Tang University of Waterloo, Waterloo, ON, Canada

Siyu Tao National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China

Yinglei Teng School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Y. Thomas Hou Virginia Tech, Blacksburg, VA, USA

Bingchuan Tian Nanjing University, Nanjing, Jiangsu, China

Rui Tian Beijing University of Technology, Beijing, China

Xiaohua Tian Shanghai Jiao Tong University, Shanghai, China

Zhi Tian ECE Department, George Mason University, Fairfax, VA, USA

Andrea M. Tonello Institute of Networked and Embedded Systems, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Fei Tong The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

Chi-Wei Tseng Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

Masahiro Umehira Ibaraki University, Hitachi, Japan

Ramachandran Venkatesan Faculty of Engineering and Applied Science, Memorial University, St. John's, NL, Canada

Vaidehi Vijayakumar School of Computing Science and Engineering, VIT University, Chennai, India

Liangtian Wan Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian, China

Shaohua Wan School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, Hubei, China

Anxi Wang School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China

Bichai Wang Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China

Chen Wang School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China

Cheng-Xiang Wang Southeast University, Nanjing, China

Chenmeng Wang Chongqing University of Posts and Telecommunications, Chongqing, China

Cong Wang Department of Computer Science, City University of Hong Kong, Hong Kong, China

Danyang Wang Xidian University, Xi'an, China

Hongwei Wang National Research Center of Railway Safety Assessment, Beijing Jiaotong University, Beijing, China

Honggang Wang Department of Electrical Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA

Hao Wang Jiangxi Normal University, Nan Chang, China

Jingjing Wang Department of Electronic Engineering, Tsinghua University, Beijing, People's Republic of China

Kai Wang Harbin Institute of Technology, Weihai, P. R. China

Kun Wang Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA

Ping Wang School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore, Singapore

Department of Information and Communication Engineering, Tongji University, Shanghai, China

Ran Wang College of Intelligence and Computing, Tianjin University, Tianjin, China

Nanjing University of Aeronautics and Astronautics, Nanjing, China

Department of Information and Communication Engineering, Tongji University, Shanghai, China

Rui Wang Department of Information and Communication Engineering, Tongji University, Shanghai, China

Siyi Wang Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China

The Institute for Telecommunications Research, University of South Australia, Adelaide, Australia

Tao Wang Department of Computer Science, New Mexico State University, Las Cruces, NM, USA

Wei Wang ECE Department, University of Waterloo, Waterloo, ON, Canada
Huazhong University of Science and Technology, Wuhan, China

Xiaofei Wang School of Computer Science and Technology, Tianjin University, Jinnan, Tianjin, China

Xingwei Wang College of Computer Science and Engineering, Northeastern University, Shenyang, China

Xiaoyan Wang Ibaraki University, Hitachi, Japan

Xiaojie Wang Dalian University of Technology, Dalian, China

Xiaoliang Wang Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu Province, China

Xuehe Wang Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore

Xuyu Wang Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA

Yawei Wang The George Washington University, Washington, DC, USA

Ye Wang Harbin Institute of Technology (Shenzhen), Shenzhen, China

Yu Wang University of North Carolina at Charlotte, Charlotte, NC, USA

Yue Wang ECE Department, George Mason University, Fairfax, VA, USA

Yuyao Wang School of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China

Zhipeng Wang EIE School, Harbin Institute of Technology, Shenzhen, China

Takashi Watanabe Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

Richard Weber Statistical Laboratory, University of Cambridge, Cambridge, UK

Steven Weber Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA, USA

Hung-Yu Wei Department of Electrical Engineering and Graduate, Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan

Yifei Wei School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Chao-Kai Wen National Sun Yat-Sen University, Kaohsiung, Taiwan

Jinming Wen College of Information Science and Technology, and College of Cyber Security, Jinan University, Guangzhou, China

Miaowen Wen School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Jian Weng College of Information Science and Technology, Jinan University, Guangzhou, China

Kevin C. Weng Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, VA, USA

Wayan Wicke Institute for Digital Communications, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany

Christian Wietfeld Communication Networks Institute (CNI), Faculty of Electrical Engineering and Information Technology, TU Dortmund University, Dortmund, Germany

Vincent W. S. Wong Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada

Chengyu Wu Zhejiang Sci-Tech University, Hangzhou, China

Di Wu School of Data and Computer Science and Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China

Fan Wu Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Hongyi Wu The Center for Cybersecurity, Old Dominion University, Norfolk, VA, USA

Qihui Wu Department of Electronics and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Qingtao Wu School of Information Engineering, Henan University of Science and Technology, Luoyang, China

Xiugang Wu Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Yuan Wu College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China

Bin Xia Department of Electronics Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China

Qiufen Xia International School of Information Science and Engineering, Dalian University of Technology, Jinzhou Area, Dalian, China

Xiang-Gen Xia Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Yili Xia Southeast University, Nanjing, China

He Xiao Department of Software Engineering, Jiangxi University of Science and Technology, Nanchang, Jiangxi, China

Liang Xiao Department of Communication Engineering, Xiamen University, Xiamen, China

Ming Xiao Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Yanping Xiao National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China

Yue Xiao National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China

An Xie Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu Province, China

Junfeng Xie School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Tian Xie Department of Electronic Engineering, Tsinghua University, Beijing, China

Xiong Xiong Beijing University of Posts and Telecommunications, Beijing, China

Zehui Xiong School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore, Singapore

Cheng Xu Hong Kong Baptist University, Kowloon Tong, Hong Kong

Fangmin Xu Beijing University of Posts and Telecommunications, Beijing, China

Guangquan Xu College of Intelligence and Computing, Tianjin University, Tianjin, China

Hansong Xu Department of Computer Engineering, Towson University, Towson, MD, USA

Jianliang Xu Hong Kong Baptist University, Kowloon Tong, Hong Kong

Lei Xu Nanjing University of Science and Technology, Nanjing, China

Mingwei Xu Tsinghua University, Beijing, China

Rongtao Xu State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Shuang Xu College of Computer Science and Engineering, Northeastern University, Shenyang, China

Shaoyi Xu Department of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

Wei Xu Southeast University, Nanjing, China

Weiqiang Xu Zhejiang Sci-Tech University, Hangzhou, China

Xiaolong Xu Nanjing University of Information Science and Technology, Nanjing, Jiang Su, P. R. China

Yiling Xu School of Computer Science and Engineering, Southeast University, Nanjing, China

Zichuan Xu School of Software, Dalian University of Technology, Jinzhou Area, Dalian, China

Qing Xue Southwest Jiaotong University, Chengdu, China

Hao Yan Shanghai Jiao Tong University, Shanghai, China

Yan Yan University of Chinese Academy of Sciences, Beijing, China

Zhiwei Yan CNNIC, Beijing, China

Bin Yang Huazhong University of Science and Technology, Wuhan, China

Chao Yang Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA

Fan Yang Beijing University of Posts and Telecommunications, Beijing, China

Guang Yang KTH Royal Institute of Technology, Stockholm, Sweden

Haojun Yang Intelligent Computing and Communications (IC²) Lab, Wireless Signal Processing and Networks (WSPN) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Jian Yang University of Science and Technology of China, Hefei, China

Jie Yang National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Jin Yang Verizon Communications, Walnut Creek, CA, USA

Kan Yang Department of Computer Science, The University of Memphis, Memphis, TN, USA

Lei Yang University of Nevada Reno, Reno, NV, USA

Liuqing Yang Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA

Shun-Ren Yang Department of Computer Science and Information Engineering, National Tsing Hua University, Hsinchu, Taiwan, Republic of China

Xiaowei Yang College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

Yi Yang Harbin Institute of Technology (Shenzhen), Shenzhen, China

Yixian Yang School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China

Zhe Yang Intelligent Computing and Communications (IC²) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Lin Yao International School of Information Science and Engineering, Dalian University of Technology, Jinzhou Area, Dalian, China

Xin-Wei Yao Zhejiang University of Technology, Hangzhou, China

Geoffrey Ye Li School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Qiang Ye Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

Department of Electrical and Computer Engineering and Technology, Minnesota State University, Mankato, MN, USA

En-Hau Yeh Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

H. Birkan Yilmaz Department of Telematics Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain

Li You National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Saleh Yousefi Department of Computer Engineering, Urmia University, Urmia, Iran

Richard Fei Yu Carleton University, Ottawa, Canada

Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Chengxiao Yu Beijing Jiaotong University, Beijing, China

Junlin Yu Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China

Nan Yu Nanjing University, Nanjing, Jiangsu, China

Shui Yu University of Technology Sydney, Ultimo, NSW, Australia

Qi-Yue Yu School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, Hei Longjiang, China

Quan Yu School of Electronic and Information Engineering, Beihang University, Beijing, China

Wei Yu Department of Computer Engineering, Towson University, Towson, MD, USA

Xingliang Yuan Faculty of Information Technology, Monash University, Melbourne, VIC, Australia

Yi Yuan School of Computing and Communications, Lancaster University, Lancaster, UK

Andrea Zanella University of Padova, Padua, Italy

Sherali Zeadally College of Communication and Information, University of Kentucky, Lexington, KY, USA

Deze Zeng School of Computer Science, China University of Geosciences, Wuhan, China

Huacheng Zeng University of Louisville, Louisville, KY, USA

Kai Zeng George Mason University, Fairfax, VA, USA

Li Zeng Jiangxi Normal University, Nan Chang, China

Xianjiao Zeng College of Intelligence and Computing, Tianjin University, Tianjin, China

Xiangping Zhai Nanjing University of Aeronautics and Astronautics, Nanjing, China

Baoxian Zhang University of Chinese Academy of Sciences, Beijing, China

Bo Zhang School of Software and Applications Technologies, Zhengzhou University, Zhengzhou, China

Ce Zhang Hong Kong Baptist University, Kowloon Tong, Hong Kong

Chen Zhang Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

The George Washington University, Washington, DC, USA

Southeast University, Nanjing, China

Dajun Zhang Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Dengyin Zhang School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China

Deyu Zhang School of Software at Central South University, Changsha, China

Haijun Zhang Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services, University of Science and Technology Beijing, Beijing, China

Hang Zhang Huawei Technologies Canada Co., Ltd., Ottawa, Canada

Heli Zhang School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, China

Hongke Zhang Beijing Jiaotong University, Beijing, China

Jianhua Zhang Beijing University of Posts and Telecommunications, Beijing, China

Jianjun Zhang Southeast University, Nanjing, China

Jun Zhang Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Sai Kung, Hong Kong

Junshan Zhang Arizona State University, Tempe, AZ, USA

Leo Yu Zhang School of Information Technology, Deakin University, Geelong, VIC, Australia

Linyuan Zhang Jiangnan Institute of Computing Technology, Wuxi, China

Mengshu Zhang School of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China

Mingchuan Zhang School of Information Engineering, Henan University of Science and Technology, Luoyang, China

Ning Zhang Peng Cheng Laboratory, Shenzhen, China

Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

College of Intelligence and Computing, Tianjin University, Tianjin, China

Qinyu Zhang School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China

Qian Zhang Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, People's Republic of China

Ruonan Zhang Northwestern Polytechnical University, Shaanxi, China

Shigeng Zhang School of Computer Science and Engineering, Central South University, Changsha, P. R. China

Weile Zhang Department of Information and communication Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China

Xinggong Zhang WangXuan Institute of Computer Technology, Peking University, Beijing, China

Xuejun Zhang School of Electronic and Information Engineering, Beijing Laboratory for General Aviation Technology (Beihang University), Beijing, China

Yan Zhang Department of Engineering and Applied Science, Memorial University of Newfoundland, St John's, NL, Canada

Tianjin Chengjian University, Tianjin, China

Yang Zhang Wuhan University of Technology, Wuhan, China

Yanhua Zhang Faculty of Information Technology, Beijing University of Technology, Beijing, China

Yiwen Zhang Anhui University, Hefei, Anhui, P. R. China

Yongmin Zhang University of Victoria, Victoria, BC, Canada

Yuan Zhang University of Electronic Science and Technology of China, Chengdu, China

Yue Zhang College of Information Science and Technology, Jinan University, Guangzhou, China

Zekun Zhang Utah State University, Logan, UT, USA

Chengcheng Zhao College of Control Science and Engineering, Zhejiang University, Hangzhou, China

Honglin Zhao Department of Communication Engineering, Harbin Institute of Technology, Harbin, China

Jian Zhao School of Electronic Science and Engineering, Nanjing University, Nanjing, China

Liang Zhao School of Computer Science, Shenyang Aerospace University, Shenyang, China

Long Zhao BUPT, Beijing, China

Quanxin Zhao Department of Technology Center, Sichuan Netop Telecom Co., Ltd., Mianyang, Sichuan, China

Wang Zhen Dalian Neusoft University of Information, Dalian Maritime University, Dalian, China

Kan Zheng Intelligent Computing and Communications (IC²) Lab, Wireless Signal Processing and Networks (WSPN) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Liang Zheng Princeton University, Princeton, NJ, USA

Yifeng Zheng Department of Computer Science, City University of Hong Kong, Hong Kong, China

Zijie Zheng School of Electronics Engineering and Computer Science, Peking University, Beijing, China

Sheng Zhong Nanjing University, Nanjing, China

Gang Zhou Computer Science, College of William and Mary, Williamsburg, VA, USA

Haibo Zhou Nanjing University, Nanjing, China

Hao Zhou School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

Pei Zhou Southwest Jiaotong University, Chengdu, China

Sheng Zhou Tsinghua University, Beijing, China

Tianqi Zhou School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China

Wanlei Zhou University of Technology Sydney, Ultimo, NSW, Australia

Wenjuan Zhou College of Intelligence and Computing, Tianjin University, Tianjin, China

Yuchen Zhou School of Telecommunications Engineering, Xidian University, Shaanxi, China

Yu Zhou Southern University of Science and Technology, Shenzhen, China
The Hong Kong University of Science and Technology, Hong Kong, China

Yong Zhou School of Information Science and Technology, ShanghaiTech University, Shanghai, China

Junlong Zhu School of Information Engineering, Henan University of Science and Technology, Luoyang, China

Kun Zhu Nanjing University of Aeronautics and Astronautics, Nanjing, China

Li Zhu Beijing Jiaotong University, Beijing, China

Lina Zhu School of Electronic and Information Engineering, Shenyuan Honors College of Beihang University, Beijing, China

E&CE Department, Illinois Institute of Technology, Chicago, IL, USA

Yanmin Zhu Shanghai Jiao Tong University, Shanghai, China

Zhengyu Zhu Zhengzhou University, Zhengzhou, China

Lei Zhuang School of Data and Computer Science and Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China

Peng Zhuang Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

Weihua Zhuang Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

Cliff Zou University of Central Florida, Orlando, FL, USA

Deyue Zou Dalian University of Technology, Dalian, China

A Securing Routing Scheme for Vehicular Networks with Cognitive Radios

► [Secure Routing in Cognitive Radio Vehicular Ad Hoc Networks](#)

Access Control

Kan Yang
Department of Computer Science, The
University of Memphis, Memphis, TN, USA

Key Applications

Any data or system applications

Definition

In the field of information security, access control (AC) is the selective restriction of access to a resource (e.g., data, system, application, etc.).

Access Control Models

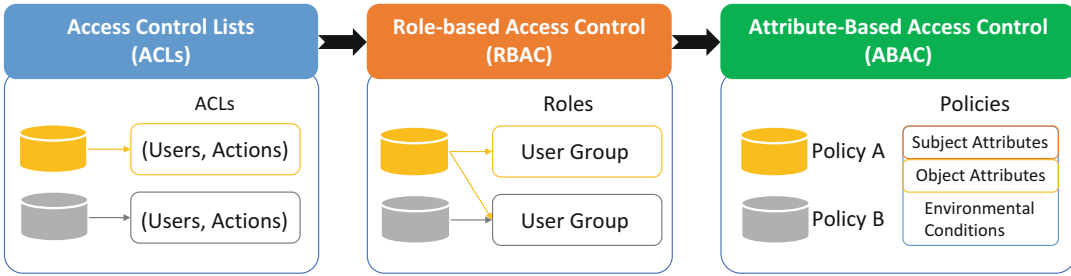
Access control is an effective approach to guarantee confidentiality and privacy of the resource.

Figure 1 shows the history of access control models: access control lists, role-based access control, and attribute-based access control.

Access Control Lists

Access control lists (ACLs) are the most basic model of access control, where each data is associated a list of mappings between the set of entities and the set of actions that each entity can take on the data. For example, the home address can be accessed by operators (read only), markets (read only), and utility service providers (read, modify, and delete). Some grid status can only be accessed by power control centers. Whenever a user tries to perform any action (e.g., modify) on the data (e.g., home address), the system or server checks the ACL of this data and determines whether to allow this action for that user. The ACLs can also be applied for a group of data (e.g., account information: name, home address, contact information), which may have the same mappings between actions and entities.

The major limitation of ACL model is that every user is treated as a distinct entity with distinct sets of privileges for each data. That means ACLs have to be defined separately for each data (or group of data), which would be a cumbersome process when many users have different levels of access permissions to a large amount of data. It is time-consuming and error-prone to selectively add, delete, and change ACLs on individual data or even group of data.



Access Control, Fig. 1 Access control models (Yang et al. 2016)

Role-Based Access Control

Role-based access control (RBAC) (Ferraiolo et al. 2001; Sandhu et al. 1996) determines whether access will be granted or denied based on user's role or function. Since many people may have the same role (e.g., market researcher), RBAC may group them into one category with a particular role, which means that the access control permission on a particular data can be set only once for all members in the group with the same role. Users can also be members of multiple groups which may have different access permissions, where more restrictive permissions override general permissions in RBAC.

Although RBAC has many advantages compared with ACL model, it still suffers from some limitations. One of the most important is that it is difficult to achieve the fine-grained access control for each user when grouping them into categories based on roles. So, how to differentiate individual members of a group is a challenging problem in RBAC.

Attribute-Based Access Control

In order to define specific access permissions for individual users, attribute-based access control (ABAC) is proposed to control data access based on the attributes associated with the user. According to the definition of NIST (Hu et al. 2014), in ABAC, the access policy is defined based on both subject attributes and object attributes under some environmental conditions. ABAC is more flexible in defining access policies than RBAC.

A key advantage of ABAC model is that no specific users are needed to be known in advance when defining access policies. The access can be granted as long as the attributes of users can satisfy access policies. Thus, ABAC is useful for the applications where organizations or data owners allow unanticipated users to be able to access the data if their attributes can satisfy some policies. This feature makes ABAC well suitable for large enterprises. An ABAC system can implement existing role-based access control policies and can support a migration from role-based to a more granular access policy based on many different attributes of individual users. Gartner predicts that "By 2020, 70% of all businesses will use ABAC as the dominant mechanism to protect critical assets, up from 5% today (in 2013)."

Encryption-Based Implementation

The ACLs, RBAC, and ABAC all require the system/server to evaluate access rules/policies and make access decisions. When the server is not fully trusted by data owners, e.g., cloud servers, how to protect data confidentiality and privacy becomes a challenging problem. In order to cope with this challenge, cryptographic techniques are applied to implement the abovementioned three access control models. Data owners encrypt files by using the symmetric encryption approach with content keys and then use every user's public key to encrypt the content keys. However, traditional public key encryption methods may pro-

duce many copies of ciphertexts for multiple users, which is not efficient in large-scale systems, e.g., smart grid. Moreover, the key management is also a complex issue in public key infrastructures (PKI).

To eliminate the need for distributing public keys (or maintaining a certificate directory), identity-based encryption (IBE) (Boneh and Franklin 2001) is a new cryptographic primitive that allows data owners to encrypt their data with identities of users instead of their public keys. To make access policies more flexible and expressive, attribute-based encryption (ABE) (Goyal et al. 2006; Bethencourt et al. 2007) is proposed for access control of encrypted data. There are two complimentary forms of ABE, namely, key-policy ABE (KP-ABE) (Goyal et al. 2006) and ciphertext-policy ABE (CP-ABE) (Bethencourt et al. 2007). In KP-ABE systems, keys are associated with access policies, and ciphertext is associated with a set of attributes, while in CP-ABE systems, keys are associated with a set of attributes, and ciphertext is associated with access policies.

Cross-References

- [Cloud Computing](#)

References

- Bethencourt J, Sahai A, Waters B (2007) Ciphertext-policy attribute-based encryption. In: Proceedings of S&P'07, Washington, DC. IEEE Computer Society, pp 321–334
- Boneh D, Franklin MK (2001) Identity-based encryption from the weil pairing. In: Proceedings of CRYPTO'01, London. Springer, pp 213–229
- Ferraiolo DF, Sandhu R, Gavrila S, Richard Kuhn D, Chandramouli R (2001) Proposed NIST standard for role-based access control. *ACM Trans Inf Syst Secur (TISSEC)* 4(3):224–274
- Goyal V, Pandey O, Sahai A, Waters B (2006) Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of CCS'06, New York. ACM, pp 89–98
- Hu VC, Ferraiolo D, Kuhn R, Schnitzer A, Sandlin K, Miller R, Scarfone K (2014) Guide to attribute based access control (abac) definition and considerations. NIST Spec Publ 800:162
- Sandhu RS, Coyne EJ, Feinstein HL, Youman CE (1996) Role-based access control models. *Computer* 29(2):38–47
- Yang K, Jia X and Shen X (2016) Privacy-preserving Data Access Control in the Smart Grid. *Cyber Security for Industrial Control Systems: from the viewpoint of close-loop*. Peng Cheng, Heng Zhang and Jiming Chen (Eds.), CRC

Acoustic Networks

- [Opportunistic Routing in Underwater Sensor Networks](#)

Acoustic Sensor Networks

- [Opportunistic Routing in Underwater Sensor Networks](#)

Acoustic Wireless Sensor Networks

- [Opportunistic Routing in Underwater Sensor Networks](#)

Ad Hoc Networks

- [Capacity of Wireless Ad Hoc Networks](#)

Adaptive Channel Access Control

- [Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs](#)

Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs

Qiang Ye^{1,2} and Weihua Zhuang¹

¹Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

²Department of Electrical and Computer Engineering and Technology, Minnesota State University, Mankato, MN, USA

Synonyms

Adaptive channel access control; Contention-Based and Contention-Free Channel Access; Distributed transmission coordination; QoS-aware hybrid medium access control (MAC)

Definitions

Adaptive medium access control (MAC) for Internet-of-Things (IoT)-based ad hoc networking refers to a distributed (infrastructure-less) link-layer mechanism to coordinate channel access for data transmissions from a varying number of IoT devices, generating either delay-sensitive services or data-hungry applications. The adaptive MAC should achieve and maintain high performance for heterogeneous services with differentiated quality-of-service (QoS) requirements (e.g., delay and throughput) in network traffic load dynamics.

Background and Key Applications

Next-generation wireless networks are envisioned to interconnect a proliferation of Internet-of-Things (IoT) devices (e.g., smartphones, smart sensors and actuators, home appliances) to achieve seamless communication interaction and diversified service customization, such as smart cities, industrial automation, and intelligent

transportation (Gubbi et al. 2013). To support an increasing number of end devices, current communication infrastructures are required to be boosted extensively with additional installation and operational cost. In the scenarios where the network infrastructures are temporarily not in place or not accessible (e.g., in hotspot areas or postdisaster areas), mobile ad hoc networking is an infrastructure-less and cost-effective networking technology to connect a large number of IoT devices via a device-to-device (D2D) communication mode (Nishiyama et al. 2015). In mobile ad hoc networks (MANETs), nodes are self-organized and initiate transmission requests over a common wireless channel (with 20MHz bandwidth at 2.4GHz or 5.9GHz spectrum frequencies) in a distributed way. To achieve consistently high link-layer performance (low packet transmission delay and high data throughput), an efficient medium access control (MAC) mechanism is required to coordinate channel access from a group of end devices by adapting to the variation of network traffic load (Natkaniec et al. 2013).

Without relying on centralized transmission coordination and device synchronization, the carrier sensing multiple access with collision avoidance (CSMA/CA)-based IEEE 802.11 distributed coordination function (DCF) (Bianchi 2000) is the most commonly used MAC scheme in existing MANETs, where packet transmissions from each node are initiated based on channel contention with an exponential backoff mechanism employed for collision avoidance. The contention-based DCF has the advantage of simplified implementation and high channel utilization in low traffic load conditions (i.e., the number of devices is low). However, its performance degrades substantially when the number of nodes becomes high, since a large portion of channel time is wasted in collided packet transmissions and collision resolution. Distributed time division multiple access (TDMA) schemes (Kanzaki et al. 2003; Wilson et al. 1993) eliminate transmission collisions by allocating time slot(s) exclusively for each device and thus achieve high resource utilization in high network load

conditions. However, the control information exchange among devices for distributed time slot acquisition makes the performance of TDMA inferior to DCF in a low network condition. Due to the performance trade-off between contention-based MAC and reservation-based time slot allocation, hybrid MAC schemes are proposed to combine the advantages of both types of MAC schemes by switching between contention-based and contention-free MAC frame structures, either periodically (Zhang et al. 2010, 2011) or adaptively based on instantaneous network load conditions (Hu et al. 2011). For most existing hybrid MAC schemes, the MAC switching decision is made upon measurement of some MAC parameters (e.g., number of idle TDMA slots (Ahmed 1997), number of lost acknowledgments (ACKs) (Hu et al. 2011), node buffer occupancy (Doerr et al. 2005)) which reflect the current network load condition. However, it is difficult to establish an analytical model between those measurement parameters and certain performance metrics (e.g., network throughput and packet delay), thus making the MAC switching decisions not optimal.

To satisfy differentiated quality-of-service (QoS) requirements from heterogeneous IoT services, the distributed MAC is expected to be both context-aware and service-aware (or QoS-aware) in a changing network environment. For example, delay-sensitive voice communications and remote control applications have stringent delay bound requirements on each transmitted packet so that the packets that are received beyond the delay bound are dropped, whereas the smart sensing applications collect data from a massive number of end devices which are throughput-oriented. Service prioritized contention schemes (e.g., busy tone-based MAC (Wang et al. 2007)) give guaranteed channel access opportunities to voice devices to relieve their contention collisions with data devices, which, on the other hand, suppress the data traffic channel access probability. Moreover, packet collisions among voice devices exist and are accumulated with the increase of voice device number. Existing distributed TDMA mechanisms can alleviate channel contention collisions but

are not suitable for supporting heterogeneous services. Since most of the data-hungry sensing applications generate event-driven data traffic, the allocated time slots can be underutilized and need to be adaptive to traffic burstiness. To meet differentiated QoS demands, an adaptive and QoS-aware hybrid MAC scheme is required to differentiate the channel access between delay-sensitive applications and high data rate applications, where time slots are allocated to delay-sensitive traffic and data nodes occupy portions of channel time via contention to exploit traffic multiplexing gain (Zhang et al. 2010, 2011). With a varying heterogeneous network traffic load, how to adaptively allocate time slots to delay-sensitive applications and adjust the channel access parameters among data devices to achieve consistently bounded packet loss rate and maximum data throughput needs investigation.

Adaptive MAC Solutions

As stated precedingly, adaptive and hybrid MAC solutions are desired to coordinate packet transmissions among devices in an IoT-based MANET for achieving consistently maximum network performance, by adapting to a varying number of end devices and providing differentiated QoS guarantee. In the following, the technical steps to develop adaptive MAC solutions are presented for a homogeneous service scenario (i.e., support only high-rate data service) and for a heterogeneous service scenario (i.e., support both voice and data applications), respectively.

Homogeneous Service Scenario

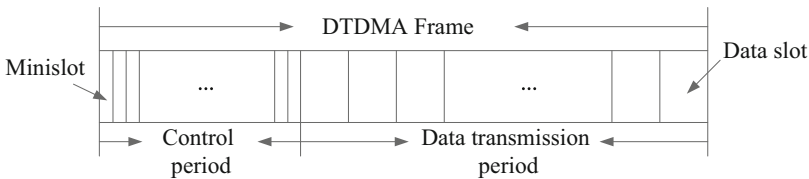
Consider a fully connected MANET (i.e., each device is within the one-hop transmission range of any other device.) with a single and error-free wireless channel. Note that developing an adaptive MAC solution for a multi-hop MANET is considered in one of our recent works Ye and Zhuang (2017a). There is no central controller in the network, and devices coordinate their packet transmissions in a distributed manner.

Packet transmission failures are due to channel contention collisions. All devices in the network are assumed homogeneous, generating the same type of high-rate data application. Each device randomly chooses its destination among other devices and can send packets to and receive packets from other devices in the half-duplex mode. Every device has a unique device identifier (ID) that is randomly selected and included in its transmitted packet headers. Devices are synchronized in time by receiving an 1PPS signal periodically with a global positioning system (GPS) receiver. The total number of devices in the network is denoted by N , which varies slowly when devices move in or move out the network coverage area. Traffic arrivals at each device are assumed to follow a Poisson process with the rate parameter λ packet/s.

To establish an adaptive MAC framework in the presence of a varying number of devices, a hybrid MAC solution (Ye et al. 2016) is proposed to switch MAC frame structures between the CSMA/CA-based IEEE 802.11 DCF and a dynamic TDMA (DTDMA) scheme (Wilson et al. 1993) under different network load conditions. The DCF with exponential backoff for collision resolution achieves high data throughput when the number of devices is low, but experiences low channel utilization as the transmission collisions are accumulated in high network load conditions. For the DTDMA, time is divided into a sequence of frames, as shown in Fig. 1. Each frame consists of a control period and a data transmission period. The control period has a number of constant-duration minislots used for local information exchange to allocate time slots in the data transmission period to each device. The time slot allocation is conducted in a distributed way to fit the MANET

scenario. The number of minislots indicates the maximum number of end devices that can be admitted in the network. The data transmission period is composed of a number of data slots, which equals the current device number in the network, and the duration of each data slot is set the same as one data packet length. We assume that all data packets have an identical and fixed packet length. Since the DTDMA eliminates packet transmission collisions, its performance is superior to the DCF in a high network load condition. Therefore, with the consideration of these two candidate MAC schemes, the proposed adaptive MAC solution uses a separate mediating MAC entity (Doerr et al. 2005) working on top of the MAC candidates maintained at each device to make MAC switching decisions according to the variation of the number of devices, N , in the network.

To maintain consistently maximum network throughput by switching between the MAC candidates, the optimal MAC switching threshold (i.e., the optimal number of devices N^*) needs to be determined. To this end, a unified and closed-form performance analytical framework (Ye et al. 2016) is established for the aggregate network throughput with respect to the number of devices N for both DCF and DTDMA schemes, based on *least-squares curve fitting* and *M/G/1 queueing analysis*. For the throughput analysis, both traffic non-saturation (i.e., the transmission buffer of each device can be empty) and traffic saturation (i.e., each device always has packets to be transmitted) conditions are considered. For a traffic non-saturation condition, closed-form throughput analytical expressions $H_1(N, \lambda)$ and $H_2(N, \lambda)$ are established in a function of N and λ for the DCF and the DTDMA, respectively. With the increase of N , packet service rate for each



Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs, Fig. 1 A DTDMA frame structure

device decreases, making the network operating in DCF or DTDMA enter the traffic saturation state. Thus, the throughput for the DCF and the DTDMA is also analyzed in a closed-form function of N , denoted by $H_3(N)$ and $H_4(N)$.

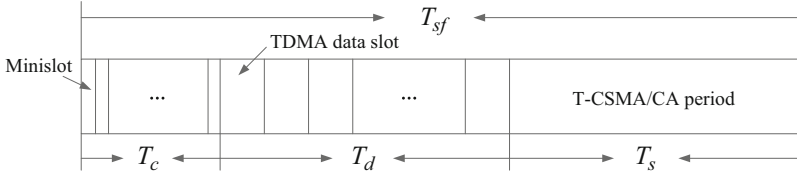
To calculate the optimal MAC switching threshold, denoted by N^* , throughput comparison between the two MAC candidates is conducted based on the developed closed-form performance analytical models. Since DCF and DTDMA have different network load points to enter the traffic saturation state, the following four combinations of traffic load states for both MAC schemes should be taken into consideration for the performance comparison: (1) the network is in the traffic saturation state for both MAC schemes; (2) the network is in a traffic non-saturation state for both MAC schemes; (3) the network is saturated for the DCF and non-saturated for the DTDMA; and (4) the network is saturated for the DTDMA and non-saturated for the DCF. Then, the optimal MAC switching threshold in terms of the number of devices, N^* , in the network can be determined, which also varies with the traffic arrival statistics λ at each device. Therefore, the adaptive MAC solution consistently makes the MAC switching decision at an optimal point (i.e., the DCF is operated when $N < N^*$ and is switched to the DTDMA when $N \geq N^*$) to achieve maximum aggregate network throughput.

Heterogeneous Service Scenario

To support heterogeneous service types, an adaptive and QoS-aware MAC solution is required. Consider the same system model as in the homogeneous service scenario except that both delay-sensitive voice devices and high-rate data devices are present in the network. The number of voice and data devices are denoted by N_v and N_d , which are slowly varying with time. For the delay-sensitive voice service, packet arrivals at each voice device follow an *on/off* model (Wang et al. 2007), which is a two-state Markov process with the *on* and *off* states being the traffic generation and traffic suppression phases, respectively. The durations each device stays in *on* and *off* states are independent and exponentially

distributed with respective average of $\frac{1}{\gamma}$ and $\frac{1}{\delta}$. During the *on* state, packets arrive periodically with the constant rate α packet/s. Each voice packet has a hard delay bound requirement, so that packets received beyond the delay bound are dropped. Therefore, the voice service has a packet loss rate bound requirement denoted by P_l . For the data service, each device is expected to access the channel by exploiting resource multiplexing gain to achieve as high as possible data throughput. All data devices are assumed in the traffic saturation state.

To satisfy the QoS requirements from both voice and data services, a QoS-aware hybrid MAC scheme is proposed to differentiate the channel access among voice and data devices (Ye and Zhuang 2017b), where voice devices are allocated time slots in a distributed way to guarantee a bounded packet delay by avoiding contention collisions and data devices contend for the channel access according to a truncated CSMA/CA (T-CSMA/CA) scheme to exploit high resource multiplexing gain. Time is partitioned into a sequence of fixed-duration superframes, as shown in Fig. 2. The duration of each superframe, denoted by T_{sf} , is set the same as the packet delay bound for voice traffic. Each superframe is composed of a control period, a TDMA period, and a T-CSMA/CA period, the durations of which are denoted by T_c , T_d , and T_s . The control period consists of a number N_{mi} of constant-duration minislots, each with an exclusive minislot sequence number (MN). Each voice device selects a unique minislot to broadcast and exchange its local information among its one-hop neighbors for distributed data time slot scheduling in the following TDMA period. Thus, N_{mi} also indicates the maximum number of voice devices (i.e., voice capacity) that can be admitted in the network. The TDMA period is divided into multiple equal-duration data transmission slots, each of which has a unique data slot sequence number (DN). Every active voice node (i.e., having non-empty transmission buffer) occupies one data slot to send a number of packets (a voice burst) generated during the previous superframe time. The number of



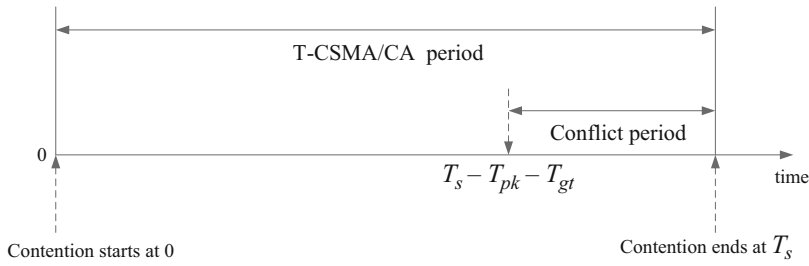
Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs, Fig. 2 A hybrid MAC superframe structure

voice bursts scheduled in each TDMA period is indicated by N_{sv} . To provide voice traffic guaranteed service quality, a maximum fraction ρ (< 1) of channel time in each superframe is allocated to voice traffic, including the control period and the TDMA period. The voice capacity N_{mi} can be determined based on ρ and the packet loss rate bound P_l . In the T-CSMA period, data devices contend the channel according to the CSMA/CA with exponential backoff with periodic interruption of the control period and the TDMA period in each superframe.

Next, a traffic-adaptive time slot allocation mechanism is presented according to instantaneous voice traffic load in the network. Each voice device randomly selects a minislot from the control period of a superframe after the time synchronization and broadcasts a control packet in the selected minislot to its one-hop neighbors. A control packet sent from a tagged device contains the following important information: (1) device IDs from its one-hop neighbors including the tagged device; (2) MN of the minislot occupied by the tagged device; (3) buffer occupancy bit (BO), $BO = 1$ if the device's transmission buffer is non-empty and 0 otherwise; and (4) previous DN, indicating the DN of the occupied data slot from the tagged device in previous superframe, with $DN = 0$ if the device was not allocated a data slot in the previous superframe. The selection of one minislot from the tagged device is considered successful if the broadcast control packets at subsequent minislots following the selected minislot contain the tagged device ID, and thus the same minislot will be selected in every subsequent superframe. Otherwise, collision happens for accessing the minislot, and all the devices involved will wait until the next superframe for

a new minislot selection. With the consideration of the *on/off* traffic statistics, only active voice devices (i.e., the devices with $BO = 1$ in broadcast control packets) are allocated data transmission time slots after accessing the minislots, to adapt to instantaneous voice traffic load. Two categories of active voice devices are defined: category I and category II. A category I device is currently active but was not allocated a data slot in previous superframe, whereas a category II device is active in both current and previous superframes. Since the activation of category I device is at some random time instant before its minislot accessing time at current superframe, category I devices are prioritized over category II devices for the time slot scheduling. Specifically, category I devices are scheduled data time slots first by following their minislot accessing sequence to minimize the packet delay bound violation probability, under the condition that each category II device can be scheduled a time slot no later than the same time slot as in previous superframe.

After the TDMA period, the data devices contend for the channel access by employing the T-CSMA/CA scheme. The T-CSMA/CA is similar as the CSMA/CA with exponential backoff, except that the contention-based packet transmissions in one superframe are periodically interrupted by the presence of its subsequent superframe. Therefore, the performance of T-CSMA/CA is different from the CSMA/CA in the following two aspects: (1) the packet waiting time before transmission is enlarged by the control period and the TDMA period of each superframe; and (2) before transmitting one packet at the end of the exponential backoff phase, each device needs to check whether the remaining time in current superframe is



Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs, Fig. 3 An illustration of T-CSMA/CA period in each superframe

enough to transmit at least one packet. If the remaining time is less than the summation of a complete packet transmission time (T_{pk}) and a guard time (T_{gt}), a virtual conflict occurs and $(T_{pk} + T_{gt})$ is the duration of the conflict period, as shown in Fig. 3. Then, all devices involved in a virtual conflict are required to hold on the packet transmissions until the beginning of subsequent T-CSMA/CA period.

The optimal parameters of the proposed hybrid MAC solution are derived, which are adaptive to varying numbers of voice and data devices to achieve bounded voice packet delay and consistently maximum data throughput. For the delay-sensitive voice service, given the maximum fraction of channel time, ρ , allocated to voice traffic in each superframe, the voice capacity N_{mi} supported in the network is analyzed, which is also the number of minislots configured for the control period of each superframe, to guarantee the voice packet loss rate bounded by P_l . Given N_v , the maximum number of voice bursts (data slots), denoted by N_{sm} , that can be scheduled in each superframe to achieve P_l is then determined. Since the actual number of scheduled data slots in each superframe is likely less than N_{sm} , the average number of scheduled data slots $\overline{N_{sv}}$ and the average duration $\overline{T_d}$ of each TDMA period are also calculated. Moreover, N_{sm} , $\overline{N_{sv}}$, and $\overline{T_d}$ are dynamically adapted to the variation of N_v to achieve a consistently bounded voice packet loss rate. After obtaining T_c and $\overline{T_d}$, the average duration $\overline{T_s}$ of the T-CSMA/CA period is also determined, which is employed for channel access by N_d data devices.

Therefore, the aggregate data throughput S_d in the T-CSMA/CA period of each superframe is further derived in terms of N_v , N_d , and the packet transmission probability τ_d from each device at a backoff slot. After some algebraic manipulation and certain approximation, the optimal transmission probability τ_d^{opt} and the corresponding optimal contention window size CW^{opt} to achieve maximum aggregate data throughput S_d^{max} are obtained in closed-form expressions of N_v and N_d . With the derived performance analytical models, the optimal MAC parameters CW^{opt} and τ_d^{opt} can also be dynamically adjusted with variations of N_v and N_d . Therefore, the proposed hybrid MAC solution in supporting heterogeneous services optimizes the MAC configurations and adapts the optimal MAC parameters to heterogeneous traffic load conditions for maintaining consistently maximum network performance.

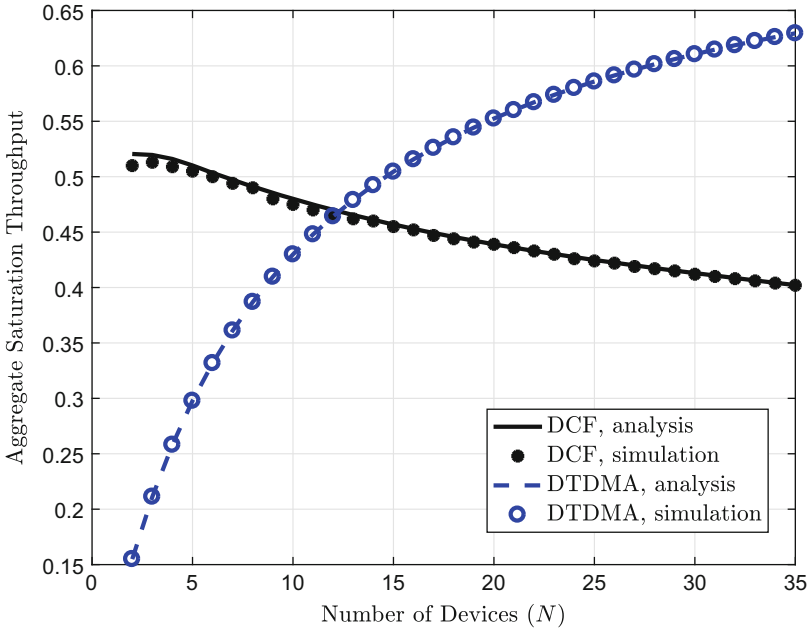
Our proposed adaptive MAC solutions for both homogeneous and heterogeneous service scenarios are useful in an IoT-based network environment, where conventional cellular communication infrastructures are inaccessible and devices are interacted via an ad hoc networking. For example, in a postdisaster area, a large number of smartphones and smart sensors are required to be interconnected for information dissemination. The proposed MAC solutions are efficient in coordinating channel access from the smart devices in an infrastructure-less manner and adapting the MAC performance to the network load fluctuations due to device activation/deactivation and device mobility.

Numerical Results

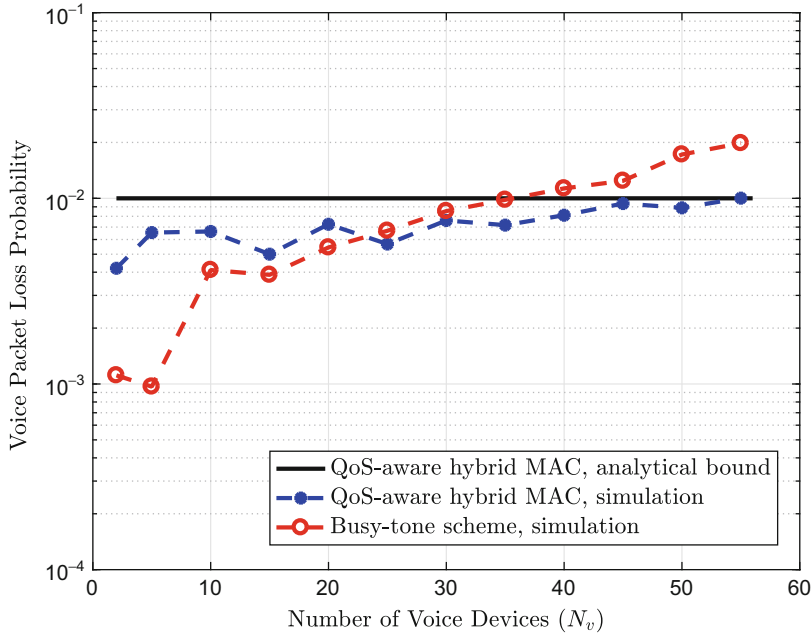
Simulation results are provided to verify the effectiveness of the proposed adaptive MAC solutions. All simulations are carried out by using the network simulator OMNeT++. In the simulation, devices are randomly scattered over a 100×100 m square region and are within one-hop communication ranges of other devices. Each source device randomly selects its destination among the other devices. For the homogeneous service scenario, devices are with the same type of high-rate data application. Packet arrivals at each data device are assumed to follow a Poisson process with the average packet arrival rate set as 300 packet/s for the traffic saturation state (results for a traffic non-saturation state are provided in Ye et al. 2016). For the heterogeneous service scenario, there are a mixture of delay-sensitive voice devices and high-rate data devices in the network. Voice traffic is generated periodically during the *on* state with the rate of 50 packet/s. Other system parameters for the simulation are provided in Ye et al. (2016) and Ye and Zhuang (2017b).

Performance comparison between DCF and DTDMA is conducted for the homogeneous service scenario. Figure 4 shows the aggregate saturation throughput for both candidate MAC protocols with the variation of N . It can be seen that the optimal MAC switching threshold, N^* , exists at the network load point when the two MAC candidates achieve the same throughput. Based on the derived closed-form performance analytical models, the optimal MAC switching threshold can be determined in a distributed way with low complexity, upon which the adaptive MAC solution achieves consistently maximum network throughput.

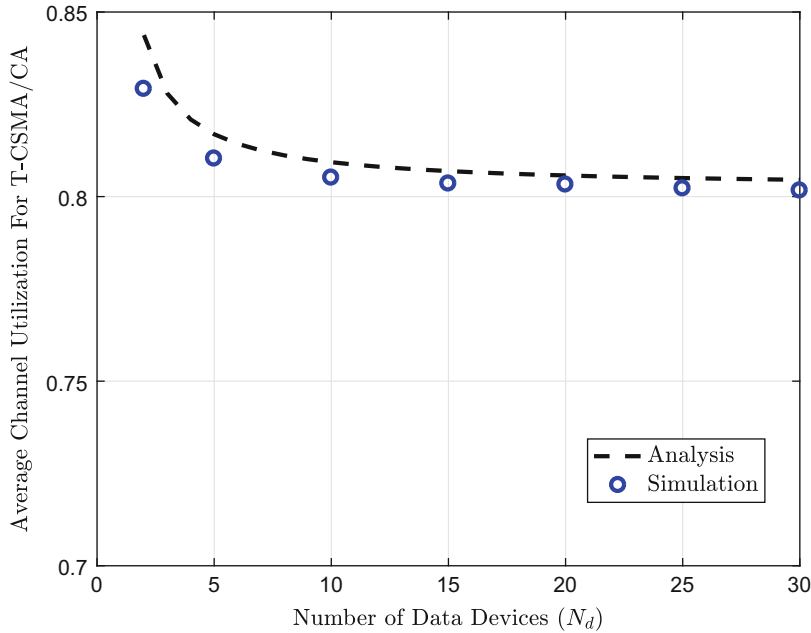
For the proposed QoS-aware hybrid MAC solution, the voice packet loss rate is evaluated in Fig. 5. It is verified through simulations that the proposed distributed and adaptive time slot allocation mechanism always guarantees a packet loss rate below the analytical bound as long as the number of admitted voice devices is within the voice capacity. Although the busy tone-based MAC scheme (Wang et al. 2007) achieves a bounded packet delay in a low network load con-



Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs, Fig. 4 Throughput comparison between DCF and DTDMA



Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs, Fig. 5 Voice packet loss rate ($N_d=10$, $\rho = 0.5$)



Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs, Fig. 6 Average channel utilization of T-CSMA/CA in each superframe ($N_v = 20$, $\rho = 0.5$)

dition, the transmission collisions among voice devices are accumulated as the network load increases, leading to degraded service perfor-

mance. The average T-CSMA/CA channel utilization within each hybrid MAC superframe is shown in Fig. 6, which is defined as the ratio of

average time for successful data transmissions inside a T-CSMA/CA period over the length of the T-CSMA/CA period. Since MAC parameters are optimized for the T-CSMA/CA, the channel utilization is consistently maximum with respect to a varying N_d .

Conclusion

In this entry, the backgrounds of distributed and adaptive MAC schemes for an IoT-enabled MANET are investigated. Novel adaptive MAC solutions are presented for both homogeneous and heterogeneous service scenarios. For the homogeneous service scenario, a hybrid MAC scheme is proposed to switch between the IEEE 802.11 DCF and the DTDMA according to the network load conditions. Based on the throughput analysis of both MAC candidates, an optimal MAC switching threshold is derived in terms of the number of devices in the network. For the heterogeneous service scenario, a QoS-aware hybrid MAC scheme is presented to differentiate the channel access for voice and data devices, where adaptive time slot allocation is conducted for voice traffic and a T-CSMA/CA scheme is employed for data traffic. The MAC parameters of the proposed scheme are optimized and are adaptive to the heterogeneous network traffic load, to achieve bound voice packet delay and consistently maximum aggregate data throughput. Simulation results are presented to verify the advantages of the proposed schemes. The applications of the adaptive MAC solutions are also discussed.

Cross-References

- [Media Access Control for Narrowband Internet of Things: A Survey](#)
- [Multiple Access Techniques](#)
- [QoS-Aware MAC](#)
- [Quality of Service in IEEE 802.11 Networks](#)

References

- Ahmed R (1997) An adaptive multiple access protocol for broadcast channels. In: Proceedings of IEEE IPCCC'97, pp 371–377
- Bianchi G (2000) Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J Select Areas Commun* 18(3):535–547
- Doerr C, Neufeld M, Fifield J, Weingart T, Sicker DC, Grunwald D (2005) MultiMAC—an adaptive MAC framework for dynamic radio networking. In: Proceedings of IEEE DySPAN'05, pp 548–555
- Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (IoT): a vision, architectural elements, and future directions. *Future Gener Comput Syst* 29(7):1645–1660
- Hu W, Yousefi'zadeh H, Li X (2011) Load adaptive MAC: a hybrid MAC protocol for MIMO SDR MANETs. *IEEE Trans Wireless Commun* 10(11):3924–3933
- Kanzaki A, Uemukai T, Hara T, Nishio S (2003) Dynamic TDMA slot assignment in ad hoc networks. In: Proceedings of IEEE AINA'03, pp 330–335
- Natkaniec M, Kosek-Szott K, Szott S, Bianchi G (2013) A survey of medium access mechanisms for providing QoS in ad-hoc networks. *IEEE Commun Surv Tutor* 15(2):592–620
- Nishiyama H, Ngo T, Oiyama S, Kato N (2015) Relay by smart device: innovative communications for efficient information sharing among vehicles and pedestrians. *IEEE Veh Technol Mag* 10(4):54–62
- OMNeT++ 5.0 [Online]. Available: <http://www.omnetpp.org/omnetpp>. Accessed on Aug. 2018
- Wang P, Jiang H, Zhuang W (2007) Capacity improvement and analysis for voice/data traffic over WLANs. *IEEE Trans Wireless Commun* 6(4):1530–1541
- Wilson N, Ganesh R, Joseph K, Raychaudhuri D (1993) Packet CDMA versus dynamic TDMA for multiple access in an integrated voice/data PCN. *IEEE J Select Areas Commun* 11(6):870–884
- Ye Q, Zhuang W (2017a) Token-based adaptive MAC for a two-hop Internet-of-Things enabled MANET. *IEEE Internet Things J* 4(5):1739–1753
- Ye Q, Zhuang W (2017b) Distributed and adaptive medium access control for Internet-of-Things-enabled mobile networks. *IEEE Internet Things J* 4(2):446–460
- Ye Q, Zhuang W, Li L, Vigneron P (2016) Traffic load adaptive medium access control for fully-connected mobile ad hoc networks. *IEEE Trans Veh Technol* 65(11):9358–9371
- Zhang R, Cai L, Pan J (2010) Performance analysis of reservation and contention-based hybrid MAC for wireless networks. In: Proceedings of IEEE ICC'10, pp 1–5
- Zhang R, Cai L, Pan J (2011) Performance study of hybrid MAC using soft reservation for wireless networks. In: Proceedings of IEEE ICC'11, pp 1–5

Adaptive Video Streaming

- [Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks](#)

Advanced Automated Meter Reading (AMR)

- [Smart Metering, Specific Challenges, and Solution Approaches](#)

Advanced Metering Infrastructure (AMI)

- [Smart Metering, Specific Challenges, and Solution Approaches](#)

Advanced Persistent Threats (APT) or Sophisticated Attacks

- [Security and Privacy in 4G/LTE Network](#)

Advances in Distribution System Monitoring

Omid Ardakanian
Department of Computing Science, University of
Alberta, Edmonton, AB, Canada

Synonyms

[Wide area monitoring system](#)

Definitions

Distribution system monitoring refers to key technologies for monitoring wide-area power distribution systems, between the substation and customer meters, to facilitate distribution grid planning and operation.

Background

Power distribution systems had a simple design historically. With radial topology, one-way power flow, and predictable demand curves, distribution system planners and operators were only required to evaluate the envelope of design conditions, such as peak loads and fault currents, to ensure reliability and power quality. Thus, there has been little need for costly telemetry beyond the substation, which was remotely monitored through supervisory control and data acquisition (SCADA) at several-second intervals. SCADA is a technology that has been in place for decades to connect sensing and control nodes, mostly in the transmission system, to a control room (Northcote-Green and Wilson 2006) using dedicated telephone lines, cellular, radio, or power line communications.

In recent years, the rapid growth in the deployment of controlled loads and distributed energy resources (DER) has led to unprecedented amounts of variability and uncertainty, which complicate distribution system planning and operation. This has necessitated more comprehensive monitoring of distribution circuits, and a novel planning and operation paradigm centered around pervasive monitoring, real-time analytics, and closed-loop control (Ardakanian et al. 2016).

Foundations

In light of new and complex grid dynamics that span multiple timescales, the operators must closely monitor the voltage and current

waveforms at different locations to detect and characterize certain behaviors, such as harmonics and oscillations, which could not be observed by the traditional SCADA system. This calls for an advanced distribution system monitoring solution which

- has low cost and complexity of deployment,
- covers both primary and secondary distribution networks,
- provides high-resolution measurements at multiple locations,
- uses precise time synchronization so that measurements can be compared across locations,
- utilizes a communication network that provides
 - stable and high bandwidth to allow for transferring high-sample-rate measurements from thousands of end nodes to upstream aggregation nodes and decentralized controllers,
 - low latency which is necessary for real-time applications, such as situational awareness, fault detection, and automated demand response,
 - high degree of reliability, or, equivalently, low link outage and packet loss probabilities,
- incorporates mechanisms to prevent unauthorized access, data manipulation, and denial of access to the sensing and control nodes,
- leverages an extensible data processing pipeline, high-throughput data stores, and real-time event triggers to provide a deeper insight into the operating state of the grid.

Two technologies have emerged in the last couple of years, namely, *smart meters* and *distribution-level phasor measurement units* (D-PMUs), which together can constitute the data acquisition layer of a well-suited distribution system monitoring system that meets the above requirements.

These technologies differ mainly in physical quantities they can measure, the time granularity of data, the location of sensors, and communication requirements. Smart meters are deployed at customer premises, recording their energy usage

and voltage level hourly or more frequently (up to every 15 min). The meters send the recorded data at regular intervals to the utility's data center using wireless or wired communications (Gungor et al. 2011). The smart meters, communications networks, and the utility's data management systems are collectively known as *advanced metering infrastructure* (AMI).

The D-PMUs (NASPI Distribution Task Team 2018) offer higher precision and resolution than smart meters. They sample voltage and current waveforms at a high frequency (usually at 120 Hz) and assign a precise time stamp to the measured quantities. The measured quantities are typically voltage and current *phasors* along with frequency, where a phasor represents the magnitude and phase angle of voltage or current. The D-PMUs are installed on distribution circuits supplementing the existing SCADA system by monitoring the network downstream of the substation. The data streams produced by a network of D-PMUs, termed *phasor network*, are aggregated by a small number of phasor data concentrators (PDCs), enabling fast comparison of time-synchronized phasor measurements from multiple locations before they reach the utility's data center. Given the bandwidth requirement of a phasor network, a broadband cellular network technology, such as 4G, is typically used for sending phasor measurements to the utility.

The smart meters and D-PMUs complement each other in the sense that they respectively monitor the end nodes and intermediate nodes in the distribution network. Moreover, in some applications, the availability of smart meter data can compensate for the lack of higher-resolution phasor measurements at some upstream nodes.

Implications for the Grid Planning and Operation

Distribution system operators can utilize the fine-grained measurements along with appropriate analytical tools for many important applications that concern planning and operation of the

Advances in Distribution System Monitoring, Table 1 Applications of distribution-level phasor measurement units

Application	Data streams	Analysis	Minimum resolution
Topology detection	Voltage phasors	On-line	1 cycle
Phase identification	Voltage phase angle	Off-line	1 s
Model parameter estimation	Voltage and current phasors	Off-line	1 s
State estimation	Voltage and current phasors	On-line	1 s
DG characterization	Voltage and current phasors	Off-line	1 min
Event detection/localization	Voltage and current phasors	On-line	1 cycle
Equipment health monitoring	Voltage and current phasors	Off-line	1 min
Outage management	Voltage and current magnitudes	On-line	1–15 min
Phasor-based control	Voltage phasors	On-line	1 cycle

distribution system (Meier et al. 2017). These applications are outlined below:

- *Topology detection* is to determine the set of energized lines and the status of switches to understand the real-time operational structure of the distribution network.
- *Phase identification* is to identify and track the connectivity and loading of the three AC phases throughout the network.
- *Model parameter estimation* is to compute impedances of distribution components, such as line segments and transformers, and update the distribution system model.
- *State estimation* (Abur and Expósito 2004) is to determine unknown state variables, for example, voltage magnitudes and phase angles at specific buses, given a set of known or measured state variables.
- *Distributed Generation (DG) characterization* is to separate the net-metered distributed generation from load.
- *Event detection and localization* is to detect and classify short-term operational and power quality events and pinpoint them to a small part of the distribution network.
- *Equipment health monitoring* is to detect and monitor equipment health issues or early signs of equipment aging to prevent damage to the equipment and support system upgrade decisions.
- *Outage management* is to automatically and reliably detect outages to reduce their duration and the system restoration cost.

- *Phasor-based control* is to incorporate phasor measurements in the feedback control of distribution system components and active end nodes, e.g., solar inverters, battery storage systems, and electric vehicle chargers.

Table 1 describes which physical quantities must be measured and what time granularity is sufficient for each of these applications.

Cross-References

- ▶ Cellular Networks: An Evolution from 1G to 4G
- ▶ Key Technologies in 4G/LTE Network
- ▶ Smart Metering, Specific Challenges, and Solution Approaches
- ▶ Power Line Communications for Grid Discovery and Diagnostics

References

Abur A, Expósito A (2004) Power system state estimation: theory and implementation. Power engineering (Willis). CRC Press, Boca Raton

Arđakanian O, Keshav S, Rosenberg C (2016) Integration of renewable generation and elastic loads into distribution grids. SpringerBriefs in electrical and computer engineering. Springer International Publishing, Cham

Gungor VC, Sahin D, Kocak T, Ergut S, Buccella C, Cecati C, Hancke GP (2011) Smart grid technologies: communication technologies and standards. IEEE Trans Ind Inf 7(4):529–539

- Meier A, Stewart E, McEachern A, Andersen M, Mehrmanesh L (2017) Precision micro-synchrophasors for distribution systems: a summary of applications. *IEEE Trans Smart Grid* 8(6): 2926–2936
- NASPI Distribution Task Team (2018) DisTT: synchrophasor monitoring for distribution systems: technical foundations and applications. Technical report, North American Synchrophasor Initiative. <https://www.naspi.org/node/688>
- Northcote-Green J, Wilson RG (2006) Control and automation of electrical power distribution Systems. CRC Press, Boca Raton

Ambient Power Technology

► Energy Harvesting Technologies in Wireless Sensor Networks

Analog and Digital Communications

Xin-Lin Huang
Tongji University, Shanghai, China

Synonyms

Analog and digital communications

Definitions

The analog signal can be one of infinite number of values, while the digital signal can only be one of finite number of values. The analog signal transmission from one part to other parts is called analog communications, while the digital signal transmission from one part to other parts is called digital communications.

Historical Background

In practice, most signals that exist in the nature are analog signals, such as voice, image, video, temperature, pressure, and so on. There are

infinite numbers of possible values for analog signal, which is continuous with time in most cases. According to the Nyquist theorem, the low-pass and band-pass analog signals can be converted into digital signals through sampling, quantization, and coding processes, which is known as analog-to-digital converting (ADC). During the ADC process, the error called quantization error is caused in the quantization process. After ADC process, the analog signal is converted into digital signal as “0” and “1” binary sequence (Hui and Yeung 2003).

The modulation modes adopted in analog communications include amplitude modulation, frequency modulation, and phase modulation. Digital communication system uses digital amplitude modulation, digital phase modulation, digital frequency modulation, and APK modulation. Since the digital signal can only be one of finite number of values, the main difference of demodulations between analog and digital communications is that a sampling and decision module is used in digital communications. The sampling and decision module maps the demodulated signals into one of the finite number of values. If the noise caused in the digital communication systems is less than the noise tolerance threshold, it can be removed by the sampling and decision unit. However, the quantization noise is irreversible in digital communications.

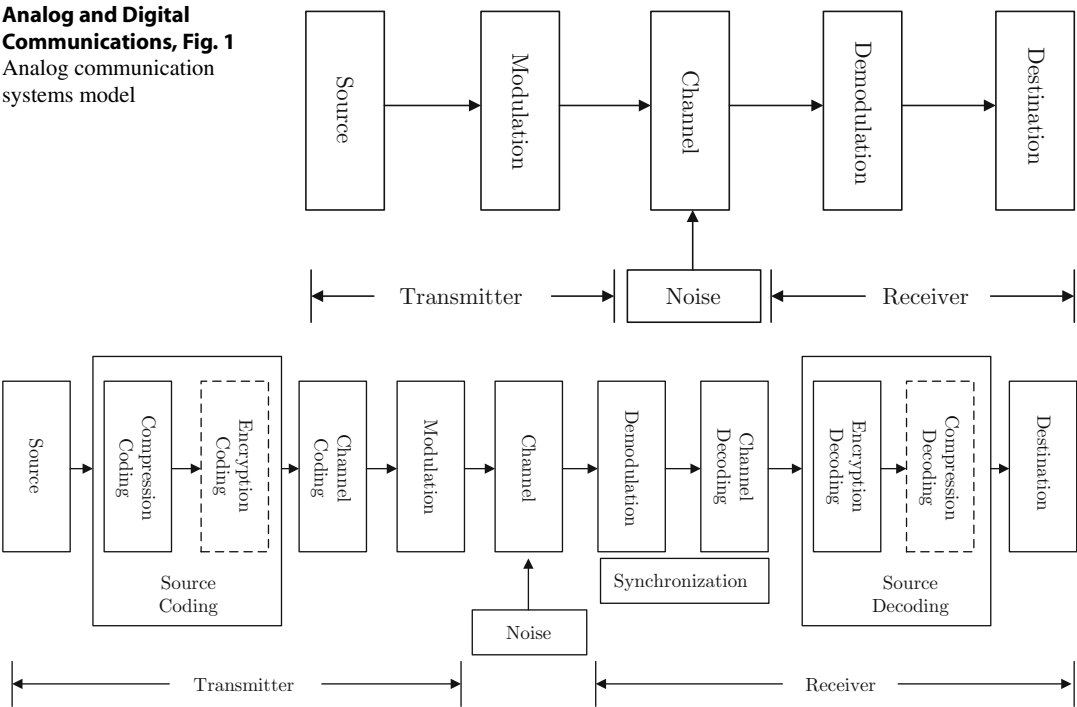
Analog and digital communications are adopted in mobile communications. The emergence of mobile communications began in the early twentieth century. The mobile communication schemes have experienced from analog communications to digital communications. In 1G to 4G communications, 1G cellular network adopts analog communications, and the others are digital communications (Kumar et al. 2010).

Foundations

Analog Communication Systems

The analog communication systems transmit analog signals, which have infinite number of possible values. Therefore, the receiver cannot com-

Analog and Digital Communications, Fig. 1
Analog communication systems model



Analog and Digital Communications, Fig. 2 Digital communication systems model

pletely eliminate the interference noise caused by the transmission channel.

Due to the amplitude distortion and phase distortion in communication systems, the performance of analog communications is affected. It means that the analog signals sent by the transmitter will not be received by the receiver accurately. The analog communications can be classified into amplitude modulation, phase modulation, and frequency modulation according to which parameter of the carrier is modulated by the transmitting analog signals. The diagram of the general analog communication systems is shown in Fig. 1. The performance of analog communication systems is merited by the signal-to-noise ratio of the reconstructed signals at the receiver.

Digital Communication Systems

The digital communication systems transmit the digital signal which can only be one of finite number of values. The sampling and decision module at the receiver can completely elimi-

nate the interference signal if it is smaller than the noise tolerance threshold and thus greatly improving the quality of digital communications and facilitating long-distance transmission with relays. The general digital communication system is shown in Fig. 2. The performance of digital communication systems is merited by the symbol error probability of the decoded digital sequence at the receiver.

Different from Fig. 1, the source coding and channel coding are adopted at the transmitter, and the corresponding source decoding and channel decoding are adopted at the receiver. Due to the overwhelming bias toward digital communications, there is no effective analog channel coding designed for analog communications.

Discussions

In digital communications, the irreversible quantization error caused by ADC process is introduced into digital transmission. The source coding leads to a serious dependence among the transmission stream and has the drawback of

error propagation during decoding. In the transmission process on time-varying fading channel, the channel quality changes dynamically. When the channel quality is higher, the digital modulation cannot obtain extra transmission quality after correctly demodulating the digital signal.

In analog communications, due to lack of effective analog channel coding scheme, the analog communications are not good at overcoming channel noise. However, analog communications can be well adapted to the dynamics of channel quality, and better signal-to-noise ratio can be obtained when the channel quality is higher. Hence, effective analog channel coding and power allocation schemes are encouraged in future research works.

Key Applications

The analog communications are mainly adopted in telephone communications, radiobroadcasting and television services, and so on. Currently, digital communications are widely used, such as computer communications, mobile communications, and digital television (Hui and Yeung 2003; Tachikawa 2003).

References

- Hui SY, Yeung KH (2003) Challenges in the migration to 4G mobile systems. *IEEE Commun Mag* 41(12):54–59
- Kumar A, Liu Y, Sengupta J, Divya JS (2010) Evolution of mobile wireless communication networks 1G to 4G. *Int J Electr Commun Technol* 1(1):68–72
- Tachikawa K (2003) A perspective on the evolution of mobile communications. *IEEE Commun Mag* 41(10):66–73

Anomaly

► Anomaly Detection for IoT Systems

Anomaly Detection for IoT Systems

Xin-Xue Lin¹, En-Hau Yeh¹, and Phone Lin²

¹Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

²Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

Synonyms

[Anomaly](#); [Internet of Things \(IoT\)](#)

Definitions

Anomaly detection for Internet of Things (IoT) system is to automatically detect whether the IoT devices, components, or systems operate normally or not. Usually there are multiple sensors or external monitors that observe the signals sent from the operating IoT systems. The detection module analyzes the signals to determine whether the system's behavior is normal or abnormal.

Historical Background

The IoT systems link the heterogeneous sensors and IoT servers to provide the IoT applications such as healthcare, industrial automation, environment monitoring, and so on. Because the decade aged IoT systems and new IoT systems may coexist, it is not easy to implement the monitors into the integrated IoT systems. It is usually to treat the integrated IoT system as a black box. Furthermore, because the signals come from one or more types of sensors (i.e., heterogeneous sensors), it is complicated for the monitors to analyze the signals from the IoT systems. One of the solutions to resolve the issue is to use the rules defined by the expert, but it usually takes time and cost. The data-driven technologies can

be applied for anomaly detection in IoT systems. Some of the previous works using these data-driven technologies are summarized as follows:

The works (Xie et al. 2017; Juvonen et al. 2015) built the statistical model based on the normal data. If a data sample does not follow the statistical model, it is judged as an abnormal data point. The work (Zhang et al. 2016) used supervised machine learning algorithms to build a classification model to check whether a data sample is normal or abnormal. It is required to label each data sample as “normal” or “abnormal” in supervised learning algorithms. The works (Valenzuela et al. 2013; Shin et al. 2011) used unsupervised machine learning algorithms to build clustering models. Because it does not rely on the labeled data, it can save lots of labeling effort.

Foundations

Anomaly detection techniques can be applied in many domains such as intrusion detection, system health monitoring, anomalous event detection in sensor networks, and so on. The anomalies could be caused by external factors (e.g., network attacking and human mistake) or internal factors (e.g., hardware failure and resource exhaustion). The anomaly detection techniques can be grouped into the following categories: density-based (Breunig et al. 2000), statistical-based (Xie et al. 2017; Juvonen et al. 2015), clustering-based (Zong et al. 2018), and the machine learning-based (Zhang et al. 2016) techniques. These techniques use historical data to build up models to determine whether the data sample is normal or abnormal. In most cases, the anomalies are rare events (i.e., less data points for the anomalies). It is very hard to directly learn the anomalous patterns from less data samples. To resolve the issue, we may only use the data samples of normal events to build up the model for anomaly detection. If a sample data does not follow the model (e.g., does not fall into the distribution of the data points of normal event), it is determined as an abnormal event.

Key Applications

In the following, we elaborate on some IoT applications (Porkodi and Bhuvaneswari 2014) for which anomaly detection can be applied:

Industrial Environment/System

In Industry 4.0 (Stojanovic et al. 2016) for automation and data exchange in the manufacturing industry, wireless sensor networks are deployed to monitor the physical environment to ensure that the manufacturing systems operate continuously and safely. With the help of the monitoring sensors, the status of the monitored object is represented by several metrics, with which the anomaly detection can determine whether the status of the monitored object is normal or abnormal.

Smart Home

In the smart home application (Fahad and Rajarajan 2015), the electrical devices, such as air-condition, television, and so on, are connected with the IoT network, through which users can interact with the devices. The anomaly detection for smart home can be applied for home security, where it can detect whether someone without authority breaks into, or check whether the door is locked or not.

Health Monitoring

With the wearable devices connected to the IoT network, the human's physiological status can be monitored remotely in real time, which makes elderly care (Shin et al. 2011) or healthcare more functional. The anomaly detection can be applied for checking whether the status deviates from distribution of the historical data, and determines whether the owner of the devices is in trouble (i.e., an emergency event occurs). It can also reduce the response time for an emergency event. Applying anomaly detection helps health monitoring system be able to react more quickly.

Connected Cars

A connected car is a vehicle with capability to connect to other vehicles, devices, and networks through wireless local area networks which assists users to drive (Kwak et al. 2016). One of the applications of the anomaly detection on the connected car is to check the user's driving behavior (e.g., it can determine whether the user has dangerous driving behavior), or check if the functionality of the car is in good status.

Smart City

In the smart city (Difallah et al. 2013), the sensors and activators (e.g., traffic light) are connected to IoT network. Examples of the smart city applications include traffic flow management and environmental monitoring. The anomaly detection in smart city can be applied to detect the car accident (i.e., abnormal traffic flow).

IoT Network Security

In the IoT network security (Hodo et al. 2016), the traditional firewall determines whether a packet can go into the Intranet by referencing the predefined security rules. However, it is hard to identify new types of attacks by using the predefined rules. The unsupervised anomaly detection mechanism can train a behavior model by using the normal Internet traffic. The model can be used to determine abnormal network traffics, e.g., attack patterns, and it can reduce the effort for human expert to find out the attack patterns.

Cross-References

- [Data-Driven Security](#)
- [Data-Driven Smart City](#)
- [Industrial Big Data](#)

References

- Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. *ACM SIGMOD Rec* 29(2):93–104
- Difallah DE, Cudre-Mauroux P, McKenna SA (2013) Scalable anomaly detection for smart city infrastructure networks. *IEEE Internet Comput* 17(6):39–47

- Fahad LG, Rajarajan M (2015) Anomalies detection in smart-home activities. In: *Proceedings of IEEE international conference on machine learning and applications*
- Hodo E, Bellekens X, Hamilton A, Dubouilh PL, Iorkyase E, Tachtatzis C, Atkinson R (2016) Threat analysis of IoT networks using artificial neural network intrusion detection system. In: *Proceedings of IEEE international symposium on networks, computers and communications*
- Juvonen A, Sipola T, Hämäläinen T (2015) Online anomaly detection using dimensionality reduction techniques for http log analysis. *Comput Netw Int J Comput Telecommun Netw* 91:46–56
- Kwak BI, Woo J, Kim HK (2016) Know your master: driver profiling-based anti-theft method. In: *Proceedings of IEEE annual conference on privacy, security and trust*
- Porkodi R, Bhuvaneswari V (2014) The internet of things applications and communication enabling technology standards: an overview. In: *Proceedings of the international conference on intelligent computing applications*
- Shin JK, Lee B, Park KS (2011) Detection of abnormal living patterns for elderly living alone using support vector data description. *IEEE Trans Inf Technol Biomed* 15(3):438–448
- Stojanovic L, Dinic M, Stojanovic N, Stojadinovic A (2016) Big-data-driven anomaly detection in industry (4.0): an approach and a case study. In: *Proceedings of IEEE international conference on big data*
- Valenzuela J, Wang J, Bissinger N (2013) Real-time intrusion detection in power system operations. *IEEE Trans Power Syst* 28(2):1052–1062
- Xie M, Hu J, Guo S, Zomaya AY (2017) Distributed segment-based anomaly detection with Kullback–Leibler divergence in wireless sensor networks. *IEEE Trans Inf Forensics Secur* 12(1):101–110
- Zhang Z, Wang X, Lin S (2016) Mobile payment anomaly detection mechanism based on information entropy. *IET Netw* 5(1):1–7
- Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho D, Chen H (2018) Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: *Proceedings of international conference on learning representations*

Anti-jamming

- [Wireless Jamming Attack](#)

Any-Path Routing

- [Opportunistic Routing \(OR\)](#)

AODV Protocol with Trust Based Mechanism

- [Trust-Based Routing in Software-Defined Vehicular Ad Hoc Networks](#)

Application of Big Data on Service Provisioning

- [Data-Driven Service Provisioning](#)

Application of Machine Learning in Wireless Sensor Network

Vaidehi Vijayakumar
School of Computing Science and Engineering,
VIT University, Chennai, India

Synonyms

[Clustering](#); [Data aggregation](#); [Machine learning](#); [Routing](#); [Wireless sensor Networks](#)

Definition

A Wireless Sensor Network (WSN) consists of spatially distributed autonomous devices using sensors to monitor physical or environmental conditions. A WSN system incorporates a gateway that provides wireless connectivity back to the wired world and distributed nodes. Machine learning (ML) is the science of getting computers to learn and act like humans do and improve their learning over time in autonomous fashion, by feeding those data and information in the form of observations and real-world interactions.

Introduction

Wireless Sensor Networks (WSN) is a vital component in Internet of Things (IoT). The small sized, low powered sensors are capable of monitoring and collecting data from environment. Most of the recent research works have not concentrated to provide a solution for analyzing the potentially huge amount of data generated by these sensor nodes. Thus, there is a need for machine learning (ML) algorithms in WSNs. When volume, velocity, and variety of data generated by WSN is very high, then data analytics tools are needed for data aggregation and clustering. ML tools are used in several applications such as intrusion detection, target tracking, healthcare, home automation, smart city. The main aim of this entry is to provide a basic knowledge in machine learning and its applications in WSN.

Background

Generally, the designers of sensor network symbolize machine learning as a branch of artificial intelligence and it is a collection of algorithms that is capable of creating prediction models. On the other hand, ML experts characterize it as a field, which is having huge amount of patterns and themes useful in sensor network applications (Alsheikh et al. 2014).

Supervised Learning

Supervised learning is merely a formulation of the concept of learning from examples. Supervised learning approach is used to resolve various issues for WSNs such as event detection, objects targeting, localization (Shareef et al. 2008), processing of query, Medium Access Control (MAC), intrusion detection, security, data integrity, and QoS.

Unsupervised Learning

No output vectors or labels are provided in unsupervised learning. Data sets are classified by finding the similarities among them. Unsupervised learning algorithm determines the concealed rela-

tionships and it can be used for solving the problems in WSN, where relationship between the variable is complex. These types of algorithms are mainly used for clustering and collection of data in WSN.

Reinforcement Learning (RL)

It allows the agent to learn from the environment by interacting with it. Here, sensor nodes learn to capture the best measurement in order to maximize the advantage. The most famous reinforcement learning algorithm is Q-learning, in which every node tries to extract measurements which are expected to increase the rewards. Sensor nodes regularly update the bonuses it has achieved derived from the actions taken at a given state. Total rewards of future can be computed by the equation:

$$\begin{aligned} Q(s_t + 1, \alpha_t + 1) \\ = Q(s_t, \alpha_t) + \gamma (r(s_t, \alpha_t) - Q(s_t, \alpha_t)) \end{aligned} \quad (1)$$

(.) indicates the incentive for taking the action a_t at state s_t and γ is the rate of learning, which decides how frequently the learning happens (between the values 0 and 1).

Machine Learning for Localization

A sensor node recognizes the position either by using GPS device or by any specific localization methods. One of such method is to collect the knowledge (e.g., connectivity, pair-wise distance measure) on the subject of the whole network into single place, where the composed information is managed in a centralized manner and identify the nodes' locations using mathematical algorithms.

Security and localization are the two main applications of Support Vector Machine in WSN. The localization method (Tran and Nguyen 2008) for mobile node uses SVM and knowledge about connectivity capabilities. By using RSSI metric (Received signal strength indicator), it detects the node location. Despite the fact that Localized SVM is capable for distributed localization in a speedy manner, it is still susceptible to outliers in training data set.

A localization system based on Self-Organizing Map (SOM) (Paladina and Paone 2007) provides a few artificial intelligence (AI) features to sensor nodes. Without any supervision, SOM can able to learn how to classify. Self-Organizing Map is used on each sensor nodes in order to evaluate the position of the node. Eight anchor nodes spatial coordinates surrounding the unspecified node form the input layer. Output layer gives the special coordinates in 2D space of the unspecified node after the training. Demerit of this method is that nodes should be organized consistently throughout the monitoring area.

An improved localization using Support Vector Regression (SVR) (Wang et al. 2016) discusses about a novel extraction method of training data for localization model. It improves the accuracy of localization method without affecting the hardware cost. The location of the unknown node is found accurately using SVM-based learning with minimum number of anchor nodes (Mary Livinsa and Jayashri 2015). Accuracy is provided by finite size of grid cells. For larger scale networks, SVM-based DV-hop is suitable.

Machine Learning for Clustering and Data Aggregation

It is really difficult to transfer entire data to sink instantly in a large scale energy-constrained sensor network. An effective solution for this problem is to send the data to an aggregator node, which is also called as cluster node. The node aggregates information from all members of the cluster and sends to the sink or base station, thus reduces the energy consumption. Numerous algorithms are actually suggested for the election of Cluster Head to increase the energy efficiency. The machine learning dependent strategies can develop the advantages of clustering and aggregation of data between nodes in WSNs through several procedures.

- ML algorithm is used to identify and remove the nonfunctional nodes from routing schemes

and compress data at CH using dimensionality reduction method.

- Machine learning algorithms are executed for electing the Cluster Head efficiently, where election of suitable cluster head will considerably decrease the usage of energy and augment the network's lifespan.

The CH election problem can be solved by the application of Decision Tree algorithm (Ahmed et al. 2010) (DT). By considering prominent features as battery, distance, and mobility, DT is electing suitable cluster head. By estimating these features, DT can provide a proficient method for identifying link reliability in Wireless Sensor Networks. Role-Free Clustering (Forster and Murphy 2010) is a clustering method with Q-learning for WSNs. The ability of the cluster head is evaluated by Q-learning algorithm with active network parameters.

A distributed spectral clustering method to group sensor nodes on their location is proposed (Muniraju et al. 2017) to avoid data congestion. This is the combination of two algorithms, distributed eigen vector computation and distributed K-Means clustering. Eigen vector of graph Laplacian is calculated by distributed power iteration technique. Clustering on eigen vector is done by distributed K-Means algorithm. In order to establish the network topology, only the location information of sensor nodes is used and this information is not exchanged in the network. In (Park et al. 2013), energy efficiency is maximized by finding cluster head with k means algorithm.

The Self-Organizing Map (SOM) is unsupervised algorithm method that learns to map from high to low dimensional space. In Cluster-based self-Organizing Data Aggregation (CODA) architecture (Lee and Chung 2006), the nodes classify the aggregated data j^* as the winning neuron, which is having a weight $W(t)$ nearest to the value of input vector $X(t)$. With CODA scheme, quality of data is increased and energy consumption and network traffic are reduced.

Principal Component Analysis (PCA) algorithm is used for dimensionality reduction and famous in the field of data compression. Principle components are a set of orthogonal variables.

Main aim is to select the significant information from data concerning the principal components. Both data compression and dimensionality reduction are multivariate methods. This can reduce the amount of data transmission between the nodes in WSN. By selecting only the momentous principal components and tossing out the lower order components, it solves the big data problem into small data. For improving the process of data aggregation, the combination of following two significant algorithms is used along with Principal Component Analysis (PCA) in WSNs (Morell et al. 2016).

- Compressive Sensing (CS): The conventional scheme of "sample then compress" is recently replaced by CS. Compressive Sensing uses sparsity property of signal.
- Expectation-Maximization (EM): This algorithm comprises of two stages, an expectation stage and maximization stage. While performing its expectation (E) stage, EM evaluates the cost function by saving the current expectation of parameters. In the maximization (M) stage, it re-computes parameters that can minimize the estimation error.

Adaptive Learning Vector Quantization (ALVQ) is used to extract compressed model of information from the nodes accurately (Lin et al. 2009). ALVQ uses the LVQ learning algorithm with previous training data for predicting the code-book. This method reduces the needed bandwidth while transmitting data and improves the accuracy of correct reading restoration from the compressed information.

Machine Learning for Routing

Considering features such as memory and computational requirements, communication costs, wireless ad-hoc nature, restricted energy, mobility and topology changes, different ML algorithms are used for energy-efficient routing in WSNs.

In fuzzy logic-based routing method (Arabi 2010), the entire network is grouped into different clusters and the election of suitable cluster head

is based on fuzzy variables. The routing in WSNs depends on the combination of hybrid routing methods and fuzzy logic for improved energy savings and improved network life span. Energy and Delay Efficient Routing protocol for sensor network (EDEAR) (Sharma and Shukla 2012) is adaptive routing method (Kumar and Kumar 2010) using Reinforcement Learning (RL). It finds the best path with minimum energy consumption and end to end delay. RL renews routing tables and thus considers all the dynamic parameters that characterize the traffic. The adaptation of routing depends on the varying traffic conditions and hence minimizes data transfer time.

Energy-aware QoS Routing algorithm using Reinforcement Learning (EQR-RL) (Jafarzadeh and Moghaddam 2014) is able to balance QoS requirements and improve network life time. Using a random load balancing algorithm, next hop neighbor is chosen. QoS parameters such as number of hops, latency, and geographical distance are considered. Dynamics of the network is learned using Q-Learning technique and routing decision is made accordingly; however, the network state information is not maintained. Sensor Intelligence Routing (SIR) using Self-Organizing Map (Barbancho et al. 2008) detects optimal routing path. The combination of SOM and Dijkstra's algorithm model offers QoS guarantees like throughput, latency, duty cycle, and packet error rate.

Example Application Scenarios

Regression-Based Adaptive Incremental Algorithm for Health Abnormality Prediction (RBAIL)

This method employs a regression based incremental algorithm for providing the learning features (Srinivasan et al. 2013). The incremental learning system is wirelessly connected to the patient and receives the stream of input parameters from patient for a fixed time of interval. RBAIL algorithm performs regression on the important health parameters for predicting the

indiscretion of patient. System uses history of the patient to check whether previous anomalies were there in order to get the updates and feedbacks. Main features here are aggregation, learning, and prediction. Correctness of the parameter is verified during aggregation. If previous data and current input data are valid and parameter value is greater than a threshold value, then abnormality is detected. If the difference of current and predicted value becomes greater than a threshold, then doctor provides feedback to correct the learning algorithm.

Object Detection and Tracking in Wireless Sensor Networks

Prediction logic is used to predict the exact location of the sink node using current location (Vaidehi et al. 2011). The estimated position is sent to Cluster Head to wake up the node which is in sleep mode. The combination of sleep wake scheduling, clustering, tracking, prediction logic, and shortest path routing minimize the energy consumption in sensor networks. Sink nodes awake cluster head that helps to reach target. Further complex events processing engine is used for detecting abnormal events in a multisensor scenario (Bhargavi and Vaidehi 2013; Gao et al. 2012).

Semantic Intrusion Detection System by Means of Pattern Matching and State Transition Analysis (SIDS)

Semantics Intrusion Detection System (SIDS) (Sri Ganesh et al. 2011) combines pattern matching, state transition, and data mining for increasing the accuracy of intrusion detection. Multiple sensors are deployed in the sensor area. The events generated by sensors are correlated in time spatial domain. The outputs from the sensors are symbolized as patterns and states. When the patterns generated by sensors violate the rule, it is detected as an intrusion. Semantics rules are developed using Another Tool for Language Recognition (ANTLR).

Online Incremental Learning Algorithm for Anomaly Detection and Prediction in Health Care

An Online Incremental Learning Algorithm (OILA) is proposed (Kirthana and Bhargavi 2014) for processing the data in online. It uses the combination of regression and feedback mechanism in order to decrease the prediction error and hence improves accuracy. The vital health parameters are received from the body sensors of a patient. Online Incremental Algorithm evaluates several parameters based on the received data and checks whether any anomalies are found. An alert is set to the doctor, if any anomalies are detected. Regression-based method is used to predict next instance. Prediction of each patient is personalized according to his/her health parameters. This algorithm calculates overall trend by *longvalue* and recent trend by *shortvalue* in health parameters. The parameters *maxthresh* and *minthresh* captures maximum and minimum threshold value of tolerance. Difference between *maxthresh* and *minthresh* is captured by a parameter *diffthresh*. Patient *sensitivity range* can be defined through sensitivity range parameter by the doctor. *History factor* is a parameter that defines number of times a patient affected to abnormalities. After reading every new instance, these parameters are updated, error is adjusted, and according to that prediction is made. The algorithm predicts the abnormality using updated parameters and triggers alert.

A Genetic Approach for Personalized Healthcare

Genetic Algorithm-based Personalized Healthcare System (GAPHS) (Vaidehi et al. 2015), uses a sensor integrated wearable chest strap for the non-invasive monitoring of physiological parameters and body parameters. Wrist wear wireless Blood Pressure (BP) sensor is used for monitoring blood pressure. A fingertip wearable oxygen saturation level (SPO2) sensor is used to detect blood oxygen saturation level. The abnormality levels of the vital parameters are classified into very low (VL), low (L), medium (M), high (H), and very high (VH) and encoded into a 5-bit

representation to determine the severity level of the patient. Using fitting function, the best chromosome that represents the personalized vital parameter of the patient is obtained. The proposed GAPHS provides an intelligent, personalized, and efficient healthcare system to serve the needy patient in right time by the doctor.

Dynamic Higher Level Learning Radial Basis Function for Healthcare Application

Traditional Radial Basis Function (RBF) has issues with using complete training set and large number of neurons. Due to these issues, computation time and complexity are increased. Dynamic Higher Level Learning RBF (DHLRBF) (Chandraskar et al. 2014) is applied to health parameters to find normal and abnormal category. The DHLRBF uses both cognitive and higher level learning component for effective classification with less complexity.

Sensor Based Decision Making Inference System for Remote Health Monitoring

Most of the existing methods have difficulty to differentiate between original and fall like patterns. Intelligent Modeling technique, Adaptive Neuro-Fuzzy Inference System (ANFIS) (Dhivya Poorani et al. 2012) is used for detecting the fall automatically with higher accuracy and less complexity. The data received from 3 axis accelerometer is categorized into five states (sit, stand, walk, lie, and fall) using ANFIS model. Mean, median, and standard deviation are selected for training the neural network. When the state is detected as fall, it examines ECG and heart rate of patient to check the abnormal condition and raise alarm.

Future Direction

The new revolution, Internet of Things is rapidly gathering momentum driven by advances in Sensor Networks and cloud technologies. The cloud has to provide security for sensed data. The storage for sensed data needs to be minimized by compression schemes. Multisensor data fusion is used for situation, object and threat refinement. The complex events sensed by sensor network

can be analyzed with data analytics tools to find the trends, patterns, and gain new insights and knowledge for variety of applications.

Cross-References

- ▶ [Data Aggregation](#)
- ▶ [Fingerprinting Localization](#)
- ▶ [Routing](#)

References

- Ahmed G, Khan NM, Khalid Z, Ramer R (2010) Cluster head selection using decision trees for wireless sensor networks. In: IEEE international conference on intelligent sensors, sensor networks and information processing, Sydney
- Alsheikh MA, Lin S, Niyata D, Tan HP (2014) Machine learning in wireless sensor networks: algorithms, strategies, and applications. *IEEE Commun Surv Tutor* 16(4):1996–2018
- Arabi Z (2010) HERF: a hybrid energy efficient routing using a fuzzy method in wireless sensor networks. In: International Conference on Intelligent and Advanced Systems (ICIAS), Manila
- Barbancho J, León C, Molina F, Barbancho A (2008) A new QoS routing algorithm based on self-organizing maps for wireless sensor networks. *Telecommun Syst* 36(2):169–185
- Bhargavi R, Vaidehi V (2013) Semantic intrusion detection with multisensor data fusion using complex event processing. *Sadhana* 38(2):169–185
- Chandraskar JB, Ganapathy K, Vaidehi V (2014) Dynamic higher level learning radial basis function for healthcare application. In: International conference on recent trends in information technology, Chennai
- Dhivya Poorani V, Ganapathy K, Vaidehi V (2012) Sensor based decision making inference system for remote health monitoring. In: International conference on recent trends in information technology, Chennai
- Forster A, Murphy AL (2010) CLIQUE: role-free clustering with Q-learning for wireless sensor networks. In: 29th IEEE international conference on distributed computing systems, Montreal
- Jafarzadeh SZ, Moghaddam MHY (2014) Design of energy-aware QoS routing algorithm in wireless sensor networks using reinforcement learning. In: 4th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad
- Kirthana R, Bhargavi VV (2014) Online incremental learning algorithm for anomaly detection and prediction in health care. In: International Conference on Recent Trends in Information Technology (ICRTIT), Chennai
- Kumar N, Kumar M (2010) Neural network based energy efficient clustering and routing in wireless sensor networks. In: First international conference on networks & communications, Chennai
- Lee SH, Chung TC (2006) Data aggregation for wireless sensor networks using self-organizing map. In: International conference on AI, simulation and planning in high autonomy systems, Berlin
- Lin S, Kalogeraki V et al (2009) Online information compression in sensor networks. In: IEEE international conference on communications, Istanbul
- Mary Livinsa Z, Jayashri S (2015) Localization with beacon based support vector machine in wireless sensor networks. In: International conference on robotics, automation, control and embedded systems (RACE), Chennai
- Mingyan Gao, Ramesh Jain et al (2012) Eventshop: from heterogeneous web streams to personalized situation detection and control. In: Proceedings of the 4th annual ACM web science conference, pp 105–108
- Morell A, Correa A et al (2016) Data aggregation and principal component analysis in WSNs. *IEEE Trans Wirel Commun* 15(6):3908–3919
- Muniraju G, Zhang S, Tepedelenlio C (2017) Location based distributed spectral clustering for wireless sensor networks. In: Sensor Signal Processing for Defense Conference (SSPD), London
- Paladina L, Paone M (2007) Self-organizing maps for distributed localization in wireless sensor networks. In: 12th IEEE symposium on computers and communications, Las Vegas
- Park GY, Kim H, Jeong HW, Youn HY (2013) A novel cluster head selection method based on K-means algorithm for energy efficient wireless sensor network. In: WAINA'14 proceedings of the 27th international conference on advanced information networking and applications, Barcelona
- Shareef A, Zhu Y, Musavi M (2008) Localization using neural networks in wireless sensor networks. In: Proceedings of the 1st international conference on mobile wireless middleware, operating systems, and applications, Turkey
- Sharma VK, Shukla SSP (2012) A tailored Q-learning/or routing in wireless sensor networks. In: 2nd IEEE international conference on parallel, distributed and grid computing, Solan
- Sri Ganesh K, Shekhar R, Vaidehi V (2011) Semantic intrusion detection system using pattern matching and state transition analysis. In: IEEE-International Conference on Recent Trends in Information Technology, ICRTIT, Chennai
- Srinivasan S, Bhargavi R, Ramkumar K, Vaidehi V (2013) An incremental algorithm technique for health abnormality prediction. In: IEEE International Conference on Recent Trends in Information Technology, ICRTIT, Chennai
- Tran D, Nguyen T (2008) Localization in wireless sensor networks based on support vector machines. *IEEE Trans Parallel Distrib Syst* 19(7):981–994

- Vaidehi V, Sandhya M, Karthika J (2011) Power optimization for object detection and tracking in wireless sensor networks. In: IEEE-International Conference on Recent Trends in Information Technology, ICRTIT, Chennai
- Vaidehi V, Ganapathy K, Raghuraman V (2015) A genetic approach for personalized healthcare. In: IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), Halifax
- Wang M, Wen-xin F, Ya-dong L, Heng-wei L (2016) An improved localization for wireless sensor network using support vector regression. In: IEEE International Conference on Computational Electromagnetics (ICCEM), Guangzhou

Applications and Architectures of Social Internet of Things

Chen Zhang^{1,2,3} and Yawei Wang²

¹Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

²The George Washington University, Washington, DC, USA

³Southeast University, Nanjing, China

Synonyms

Social internet of things (SIoT)

Historical Background

Internet of Things (IoT) has been an integral component of our lives and thriving with new products and services. This term was first introduced by Kevin Ashton at Procter & Gamble (*P&G*) in 1999 (Ashton et al. 2009). He presented the idea that computers could be empowered to observe, identify, and understand the physical world with RFID (radio-frequency identification) and sensor technology. By now, this definition has evolved continually with the development of embedded systems, wireless sensor networks, machine learning, etc. As defined in ITU (2012), IoT is the network of interconnecting physical and virtual things which

have embedded information and communication technologies that allow these things to interact and exchange data. In addition to the personal devices such as smartphones, tablets, and laptops, IoT objects have been widely used in others fields which include, but not restricted to, agriculture (Mekala and Viswanathan 2017), smart city (Arasteh et al. 2016), education (Gul et al. 2017), manufacturing (Yang et al. 2018), medical and healthcare (Joyia et al. 2017), etc.

Existing technologies have enhanced the computing capacity of IoT objects, but most of them embedded in specific and necessary hardware features are only able to finish a particular task. In Guinard et al. (2010), the authors propose a platform that enables an individual to share the services offered by his smart objects with his friends or their objects through social networks of humans. Using interconnected objects to extend the Internet and consequently leading to efficient conversations is addressed in the study (Mendes 2011). In Atzori et al. (2011), the concept of *Social Internet of Things (SIoT)*, which is a social network among the objects of IoT, is formalized. Generally, a social network for human beings is a platform (like Facebook, Twitter, Instagram, etc.) which people and organizations can share user-generated content such as videos, photos, comments, posts, location, and so on. Members of this kind of platform prefer to interact and build social relationships via the Internet with those who share similar backgrounds, interests, or friends. For objects, the SIoT is a platform for global interconnected IoT objects to obtain social-like capabilities, which allow these objects to autonomously establish social relationships, request or provide services, and collaborate with others to achieve a common goal in an autonomous way with proper authorization from users.

As a new paradigm of IoT, SIoT introduces the social relationship concept among IoT objects. This does not mean that a common architectural model of social network services for human beings can be directly applied to SIoT. The differences between main components of social network platforms for humans (Boyd and Ellison 2007) and IoT objects are presented in Atzori

et al. (2011). And the explanations of main components of social network platforms for IoT objects are organized as follows:

- *ID management (ID)*: assigns a unique ID to each IoT object and manage existing objects.
- *Object profiling (OP)*: contains static and dynamic information about the object.
- *Owner control (OC)*: sets up rules to carry out access control, relationship control, and any other operations.
- *Relationship management (RM)*: is an important component based on the user-defined rules to establish, update, and terminate relationships among objects in a network.
- *Service discovery (SD)*: finds objects that can provide required services from the network, a very similar way to searching for friends and information from an online social network.
- *Service composition (SC)*: enables interaction between objects.
- *Trustworthiness management (TM)*: provides trust evaluation among objects for relationship management and information sharing.

In a SIoT network, several kinds of relationships can be defined among objects. The types and characteristics of social relations among IoT objects are systematically summarized in Atzori et al. (2012):

- *Parental object relationship (POR)*: established when objects are created in the same period by the same manufacturer.
- *Co-location object relationship (CLOR)*: established when objects are deployed in the same location.
- *Co-work object relationship (CWOR)*: established if the objects work on the same task.
- *Ownership object relationship (OOR)*: established when objects have the same owner.
- *Social object relationship (SOR)*: established when the owners get in touch and the objects are active.

These relationships among objects in SIoT should be established and managed *without human intervention*. Several SIoT architectures

based on the above components and relationships are introduced in the next section.

SIoT Architectures

A three-layered architecture model proposed in Zheng et al. (2011) for IoT should be introduced before the SIoT architectures, since several existing SIoT architectures are built upon this IoT architecture. In this architecture model, the sensor data are acquired by sensing devices in the *Sensing Layer*. Then these data are safely and reliably transmitted to the *Application Layer* through the *Network Layer*. The *Network Layer* is devoted to transferring data across different networks such as cellular networks, WLAN (wireless local area network), satellite communication networks, and so on. By applying distributed computing technologies to massive data, the *Application Layer* performs data processing and analysis intelligently to support intelligent control in the IoT.

The following are SIoT architectures summarized from existing research:

- An architecture consisting of the Server Side and Client Side is proposed in Atzori et al. (2011). This architecture has three layers on both sides. On the Client Side, the Objects Layer consists of different physical objects. The Object Abstraction Layer is used to harmonize communications among physical objects. The top layer on the Client Side consists of the Social Agent element and the Service Management element. The Social Agent element is designed for friendships and object profile updates, service discovery, and service request from the network. Management element provides interface for humans to control the object behaviors. On the Server Side, the Base Layer provides data storage and data management. The Component Layer includes the main components of SIoT (OP, DC SD, TM, ID, RM, and SC). Application Layer consists of interfaces to human, objects, and third-party services.

- A three-layer SIoT architecture is proposed in Atzori et al. (2012). This proposed system consists of three subsystems: the SIoT Server, the Gateway, and the Object. The SIoT Server has a Network Layer and an Application Layer. The Application Layer includes three sub-layers. The database of data storage, data management, and relevant descriptors is in the Base Sub-layer. The main components of the SIoT system discussed in the previous section are located in the Component Sub-layer; they provide main capabilities for the SIoT system. The Interface Sub-layer consists of the third-party interfaces for objects, humans, and services. The Gateway and the Object both have Sensing Layer, Network Layer, and Application Layer. The characteristics of devices decide the combination of these layers. But no matter what the combination is, the Application Layer includes SIoT Applications, the Social Agent, and the Service Management. The Social Agent communicates with the SIoT server to update its profile and friendship and also provides service discovery and service request. When it is necessary, it is devoted to direct communications between objects which are close to each other in the physical world as well. The Service Management Agent provides interfaces for humans to control the behaviors of the objects.
- An architecture with Actors, the Intelligent system, the Interface, and the Internet is introduced in Ortiz et al. (2014). The Internet is the foundation of this architecture; Actors in the architecture refer to humans and things who have equal rights to take part in managing data, executing queries, and receiving services and commands. The Intelligent system is devoted to coordinating the interactions between Actors; it also has capabilities of recommendation, data and context management, service discovery and search, and service and application management. The Interface provides ports for inputs and outputs of the interactions with the system.
- A distributed sensor storage for SIoT is presented in Wu et al. (2015). The Sensing Layer is responsible for collecting data from the physical world through sensors and other

objects. The Communication Layer transfers the sensing data collected from Sensing Layers to the users via the Internet, mobile network, etc. The Application Layer consists of consumer applications, enterprise applications, and government applications. Different services can provide data for different users. In this architecture, the IoT, consisting of the Internet and sensor network, can be organized into SIoT based on the social network. In SIoT scenarios, the sensors with large storage can help normal sensors which can only sense the environment to store the sensing data. The storage should be able to repair the lost fragment and to protect data secrecy.

- A four-layered SIoT architecture is introduced in Gulati and Kaur (2019). In this architecture, the Base Layer called Object Layer is where the IoT devices and sensors are deployed. Communication and collaboration among these objects are through local sensor networks. Communication technologies and protocols used by SIoT are defined in the Communication Layer. The SIoT Management Layer defines the SIoT platforms to manage SIoT services. The main components of SIoT that were introduced in the previous section are in this layer. The top layer is the Application Layer which provides APIs (application programming interfaces) for SIoT applications.

Key Applications

Following the theoretical development of SIoT, several projects have been proposed to integrate IoT objects into a social network. Existing SIoT platforms such as *Toyota Friend Network*, *Nike+*, *Xively*, and *Paraimpu* are summarized in Atzori et al. (2014). *Toyota Friend Network* is a private social network that allows Toyota customers to create a virtual community among them and interact with their own cars. For example, customers can get a maintenance alert from their cars and can connect to the dealerships to get service information through this social network. *Nike+* provides a platform for customers to share their fitness data in a social network. *Xively*

and Paraimpu are two similar platforms which enable objects to be connected to each other or to be linked to existing social networks so that these objects can respond to their owners' requests. Furthermore, Paraimpu allows users to share the data generated by objects with their friends. A protocol for traffic management in *social Internet of Vehicles (SIOV)* (an extension to the concept of SIOt) is presented in Jain et al. (2018). This protocol can improve the performance of data transmission over vehicular social networks. Other cases that use SIOt are illustrated in Afzal et al. (2019), for instance, smart shopping mart retailing, healthcare and telemedicine, traffic surveillance and road safety, and so on. In Fu et al. (2017), an intelligent painting device intended for children is designed based on SIOt. This device aims at creating a social network consisting of humans, objects, and emotion to facilitate the children's intellectual development and personality development.

Security and Trust in SIOt

As IoT devices are becoming more ubiquitous, IoT security has been a growing concern. Several IoT security incidents make people realize the importance of IoT security. In 2016, the largest DDoS attack was launched by an IoT botnet (Goodin 2016), a malware called *Mirai* that brought people's attention to the seriousness of IoT security. Beyond that, there are even more serious attacks related to daily life. For example, the vulnerable web camera at home can be used to pry into the owner's daily life by the attacker (Whittaker 2015). A Jeep over the Internet can be hijacked, and the attacker can gain access to many functions that could be used to cause a traffic accident (Rouse 2015). In 2017, the Food and Drug Administration (FDA) warned that cybersecurity vulnerabilities are identified in several medical devices (FDA 2017).

In order to address the security threats in IoT, a series of studies are carried out from different aspects including encryption, authentication, access control, network security, and application security (Zhou et al. 2018; Alaba et al. 2017;

Diro and Chilamkurti 2018; Jia et al. 2018; Liang et al. 2018; Zheng et al. 2018). In addition, object diversity and complex relationships among objects make security communication and trustworthiness management become more important and serious for SIOt (Nitti et al. 2014). Furthermore, related policies and laws should also be formulated to regulate the ethical use of SIOt (Berman and Cerf 2017).

Conclusion

As a new paradigm of the Internet of Things (IoT), the Social Internet of Things (SIOt) will create a brand-new way of communication and interaction for IoT objects. This article systematically introduces the definition and components of SIOt as well as the existing architectures and applications. Furthermore, security and trust issues which could bring great challenges to the development of SIOt are discussed at the end of this article.

Cross-References

- [Internet of Medical Things](#)
- [Social IoT Crowd-Sourcing on Disaster Reduction](#)

References

- Afzal B, Umair M, Shah GA, Ahmed E (2019) Enabling IoT platforms for social IoT applications: vision, feature mapping, and challenges. *Futur Gener Comput Syst* 92:718–731
- Alaba FA, Othman M, Hashem IAT, Alotaibi F (2017) Internet of things security: a survey. *J Netw Comput Appl* 88:10–28
- Arasteh H, Hosseinneshad V, Loia V, Tommasetti A, Troisi O, Shafie-Khah M, Siano P (2016) IoT-based smart cities: a survey. In: 2016 IEEE 16th international conference on environment and electrical engineering (EEEIC). IEEE, pp 1–6
- Ashton K et al (2009) That “internet of things” thing. *RFID J* 22(7):97–114
- Atzori L, Iera A, Morabito G (2011) Siot: giving a social structure to the internet of things. *IEEE Commun Lett* 15(11):1193–1195

- Atzori L, Iera A, Morabito G, Nitti M (2012) The social internet of things (siot)—when social networks meet the internet of things: concept, architecture and network characterization. *Comput Netw* 56(16):3594–3608
- Atzori L, Iera A, Morabito G (2014) From “smart objects” to “social objects”: the next evolutionary step of the internet of things. *IEEE Commun Mag* 52(1):97–105
- Berman F, Cerf VG (2017) Social and ethical behavior in the internet of things. *Commun ACM* 60(2):6–7
- Boyd DM, Ellison NB (2007) Social network sites: definition, history, and scholarship. *J Comput-Mediat Commun* 13(1):210–230
- Diro AA, Chilamkurti N (2018) Distributed attack detection scheme using deep learning approach for internet of things. *Futur Gener Comput Syst* 82:761–768
- FDA (2017) Cybersecurity vulnerabilities identified in st. jude medical’s implantable cardiac devices and merlin@home transmitter: FDA safety communication. <https://www.fda.gov/medical-devices/safety-communications/cybersecurity-vulnerabilities-identified-st-jude-medicals-implantable-cardiac-devices-and-merlinhome>
- Fu Z, Lin J, Li Z, Du W, Zhang J, Ye S (2017) Intelligent painting based on social internet of things. In: *International conference on distributed, ambient, and pervasive interactions*. Springer, pp 335–346
- Goodin D (2016) Record-breaking DDoS reportedly delivered by >145k hacked cameras. *Ars Technica* 28. https://arstechnica.com/?post_type=post&p=966459 Accessed 10 February 2019
- Guinard D, Fischer M, Trifa V (2010) Sharing using social networks in a composable web of things. In: *PerCom workshops*, pp 702–707
- Gul S, Asif M, Ahmad S, Yasir M, Majid M, Malik M, Arshad S (2017) A survey on role of internet of things in education. *Int J Comput Sci Netw Secur* 17(5):159–165
- Gulati N, Kaur PD (2019) When things become friends: a semantic perspective on the social internet of things. In: *Smart innovations in communication and computational sciences*. Springer, pp 149–159
- ITU (2012) Overview of the internet of things. <http://handle.itu.int/11.1002/1000/11559>
- Jain B, Brar G, Malhotra J, Rani S, Ahmed SH (2018) A cross layer protocol for traffic management in social internet of vehicles. *Futur Gener Comput Syst* 82:707–714
- Jia Y, Xiao Y, Yu J, Cheng X, Liang Z, Wan Z (2018) A novel graph-based mechanism for identifying traffic vulnerabilities in smart home IoT. In: *IEEE INFOCOM 2018-IEEE conference on computer communications*. IEEE, pp 1493–1501
- Joyia GJ, Liaqat RM, Farooq A, Rehman S (2017) Internet of medical things (ioMT): applications, benefits and future challenges in healthcare domain. *J Commun* 12:240–247
- Liang Y, Cai Z, Yu J, Han Q, Li Y (2018) Deep learning based inference of private information using embedded sensors in smart devices. *IEEE Netw* 32(4):8–14
- Mekala MS, Viswanathan P (2017) A survey: smart agriculture IoT with cloud computing. In: *2017 international conference on microelectronic devices, circuits and systems (ICMDCS)*. IEEE, pp 1–7
- Mendes P (2011) Social-driven internet of connected objects. IAB workshop on interconnecting smart objects with the internet
- Nitti M, Girau R, Atzori L (2014) Trustworthiness management in the social internet of things. *IEEE Trans Knowl Data Eng* 26(5):1253–1266
- Ortiz AM, Hussein D, Park S, Han SN, Crespi N (2014) The cluster between internet of things and social networks: review and research challenges. *IEEE Internet Things J* 1(3):206–215
- Rouse M (2015) What is car hacking? Definition from whatis.com. <https://internetofthingsagenda.techtarget.com/definition/car-hacking>
- Whittaker Z (2015) New security flaws found in popular IoT baby monitors. <https://www.zdnet.com/article/security-vulnerability-flaw-internet-things-baby-monitors>
- Wu J, Dong M, Ota K, Liang L, Zhou Z (2015) Securing distributed storage for social internet of things using regenerating code and blom key agreement. *Peer-to-Peer Netw Appl* 8(6):1133–1142
- Yang C, Shen W, Wang X (2018) The internet of things in manufacturing: key issues and potential applications. *IEEE Syst Man Cybern Mag* 4(1):6–15
- Zheng L, Zhang H, Han W, Zhou X, He J, Zhang Z, Gu Y, Wang J et al (2011) Technologies, applications, and governance in the internet of things. *Internet of things-global technological and societal trends From smart environments and spaces to green ICT*
- Zheng X, Cai Z, Li Y (2018) Data linkage in smart internet of things systems: a consideration from a privacy perspective. *IEEE Commun Mag* 56(9):55–61
- Zhou R, Zhang X, Du X, Wang X, Yang G, Guizani M (2018) File-centric multi-key aggregate keyword searchable encryption for industrial internet of things. *IEEE Trans Ind Inform* 14(8):3648–3658

Applications of Molecular Communication Systems

Tadashi Nakano, Yutaka Okaie, and Takahiro Hara
Osaka University, Suita, Japan

Synonyms

Molecular, biological, and multiscale communications; Nano-networks

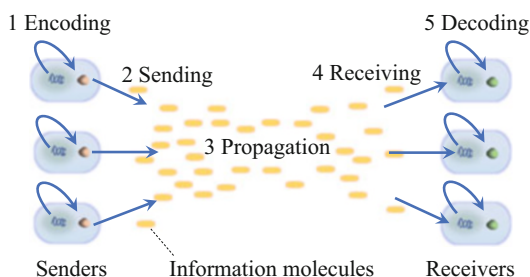
Definitions

A molecular communication system is defined as a system of bio-nanomachines that transmit and receive information using chemical signals or molecules. A bio-nanomachine that constitutes a molecular communication system is made of biomaterials with or without non-biological materials, approximately 1–100 μm in size, and capable of processing molecules. Examples of molecular communication systems are naturally occurring biological systems such as bacterial populations, epithelial sheets, and immune systems where biological cells represent bio-nanomachines. Examples of molecular communication systems also include artificial or synthetic biological systems designed for specific applications such as biomolecular sensing and targeted drug delivery.

Historical Background

Molecular communication was proposed as an unexplored research area at the intersection of communications engineering and biology (Nakano 2017). The basic concept of molecular communication is simple (Fig. 1); sender and receiver bio-nanomachines communicate using chemical signals or molecules. To transmit information, sender bio-nanomachines generate “information molecules” that represent information (i.e., encoding in Fig. 1) and transmit the information molecules into the environment (sending). Information molecules propagate in the environment (propagation) and react to receiver bio-nanomachines (receiving and decoding). Namely, receiver bio-nanomachines receive information from information molecules.

Molecular communication research can be divided into its first phase (2005–2009) and second phase (2010–present). The first phase of research (2005–2009) focuses on experimentally recreating functionalities and components of naturally occurring biological systems. This includes the first prototype implementation of molecule transport systems using motor proteins



Applications of Molecular Communication Systems, Fig. 1 Molecular communication (Nakano 2017)

and DNA molecules (Hiyama et al. 2008a). This also includes experimental demonstration of signal amplification mechanisms through calcium signaling (Nakano et al. 2009), self-organized microtubule networks (Enomoto et al. 2006), and encapsulation of information molecules and transport of information molecules from liposomes to biological cells (Moritani et al. 2010).

The second phase of molecular communication research (2010–present) has seen a wider variety of research topics concerning molecular communication. This phase of research includes the use of communication theory and mathematical tools to understand physical properties of molecular communication (Mahfuz et al. 2010; Farsad et al. 2012; Pierobon and Akyildiz 2013). It also includes design, analysis, and computer simulations of higher-layer mechanisms of molecular communication (Lio and Balasubramaniam 2012; Felicetti et al. 2014). It further includes standardization efforts to establish molecular communication as a standard framework and to develop simulation tools (Bush et al. 2015).

Key Applications

Functional applications of molecular communication systems are anticipated in a variety of domains (Nakano et al. 2012).

- **Biomolecular sensing:** Specific molecules in the human body serve as biomarkers

for certain diseases or medical conditions. More detailed information such as the spatial distribution of molecules is potentially useful for in-depth diagnosis. For biomolecular sensing, bio-nanomachines may sense their environment, communicate with others, and collectively determine whether specific molecules exist in their environment. Alternatively, bio-nanomachines may transmit sensed information to external devices or control units for diagnosis of their environment (Rogers and shung Koh 2016; Abdi et al. 2017).

- **Targeted drug delivery:** The delivery of drug molecules to target sites in the human body is expected in nanomedicine; it maximizes the efficacy of drug molecules and at the same time reduces potential side effects at nontarget sites. For drug delivery, bio-nanomachines may be embedded with drug molecules and either injected directly into target sites or intravenously to propagate to target sites. Bio-nanomachines may use molecular communication to collaborate to search for target sites, aggregate at the target sites, and release embedded drug molecules (Okaie et al. 2016; Wei et al. 2013).
- **Tissue regeneration:** In tissue development, biological cells communicate through synthesizing growth factor molecules and transmitting them into the environment. Growth factor molecules propagate in the environment and bind to cell surface receptors of the target cells. The concentrations and types of growth factor molecules modulate migration, proliferation, and differentiation of target cells, leading to the formation of a tissue structure. For the repair or construction of a tissue structure, bio-nanomachines made of living cells may be deployed in the human body. Bio-nanomachines use molecules to communicate with tissue-forming cells, while they divide and grow to help the tissue structure formation.
- **Internet of Bio-NanoThings:** Molecular communication systems may also be interfaced to and integrated with existing communication systems. Future mobile

phones or wearable devices may be integrated with bio-nanomachines capable of molecular communication for on-chip analysis of biochemical signals (e.g., molecules in blood or from sweat) (Hiyama et al. 2008b). Further, such devices and molecular communication systems may be integrated into the Internet to form the Internet of NanoThings (Akyildiz and Jornet 2010) or the Internet of Bio-NanoThings (Akyildiz et al. 2015).

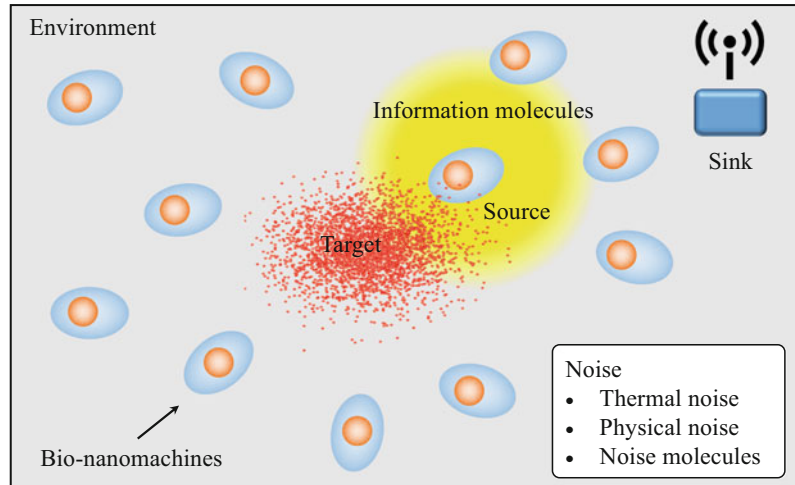
Molecular Communication System Model

Figure 2 shows a model of molecular communication systems.

- **Bio-nanomachines** are sensors and actuators to form molecular communication systems. A bio-nanomachine is characterized with the three criteria: material, size, and functionality (Nakano et al. 2012). A bio-nanomachine is composed of biomaterials with or without non-biological materials. The size of a bio-nanomachine ranges from the size of a macromolecule to that of a biological cell (i.e., up to tens of μm). A bio-nanomachine implements a set of biochemical functionalities such as sensing a certain type of molecule to achieve application-dependent goals (e.g., biomolecular sensing). Examples of bio-nanomachines are genetically engineered biological cells.
- The **environment** is a three-dimensional space where bio-nanomachines are deployed. It is typically an aqueous environment (e.g., an internal environment of the human body), containing molecules and energy sources for bio-nanomachines to operate. The environment may also contain flow by which the operation of bio-nanomachines can be disturbed or alternatively enhanced (Kadloor et al. 2012; Tavakkoli et al. 2017). The environment may generate events or information of interest, or contain targets that generate such information. For example, in biomolecular sensing applications, a change of the human body into an abnormal state

Applications of Molecular Communication Systems, Fig. 2

Molecular communication system model



represents such an event, and in targeted drug delivery applications, cancer cells represent a target.

- **Sources** are entities in the environment that provide useful information. A source may be a bio-nanomachine or a group of bio-nanomachines that detects an event or target in the environment. A bio-nanomachine functioning as a source may detect an event or target biochemically and transmit the detected information by propagating information molecules to other bio-nanomachines or sinks.
- **Sinks** are entities that receive and collect information from bio-nanomachines or sources. A sink may be a bio-nanomachine or a group of bio-nanomachines that processes information and performs application-dependent functionalities such as releasing drug molecules. A sink may also be a conventional device capable of traditional communication (e.g., wireless communication) (Nakano et al. 2014). A sink may be made from materials that are not compatible with the environment and may be orders of magnitude larger than bio-nanomachines. A sink may function as a gateway that interconnects molecular communication systems with external systems. Examples of sink devices include implantable medical devices (Kiourti et al. 2014).

- **Noise** exists in molecular communication systems in various forms. The first type of noise is thermal noise. Due to thermal noise, molecular communication among bio-nanomachines and the operation of bio-nanomachines become stochastic. The second type of noise is physical noise. The high viscosity of the environment and fluid in the environment generate physical force, making it difficult for molecules to propagate and for bio-nanomachines to move. The third type of noise is caused by molecules existing in the environment or noise molecules. Due to noise molecules, molecular communication among bio-nanomachines, and the operation of bio-nanomachines can be disturbed.

Target Detection and Tracking

Okaie et al. (2016) describes key functionalities of molecular communication systems: target detection and tracking. Target detection is a functionality of molecular communication systems to detect a target in a given environment, while target tracking is aimed at detecting and tracking targets as they move. In nanomedical applications, targets can be disease sites, pathogens, infectious microorganisms, or biochemical weapons that represent a potential threat to the environment;

the timely detection of targets and tracking of targets are important to provide immediate treatments or further analysis of the environment.

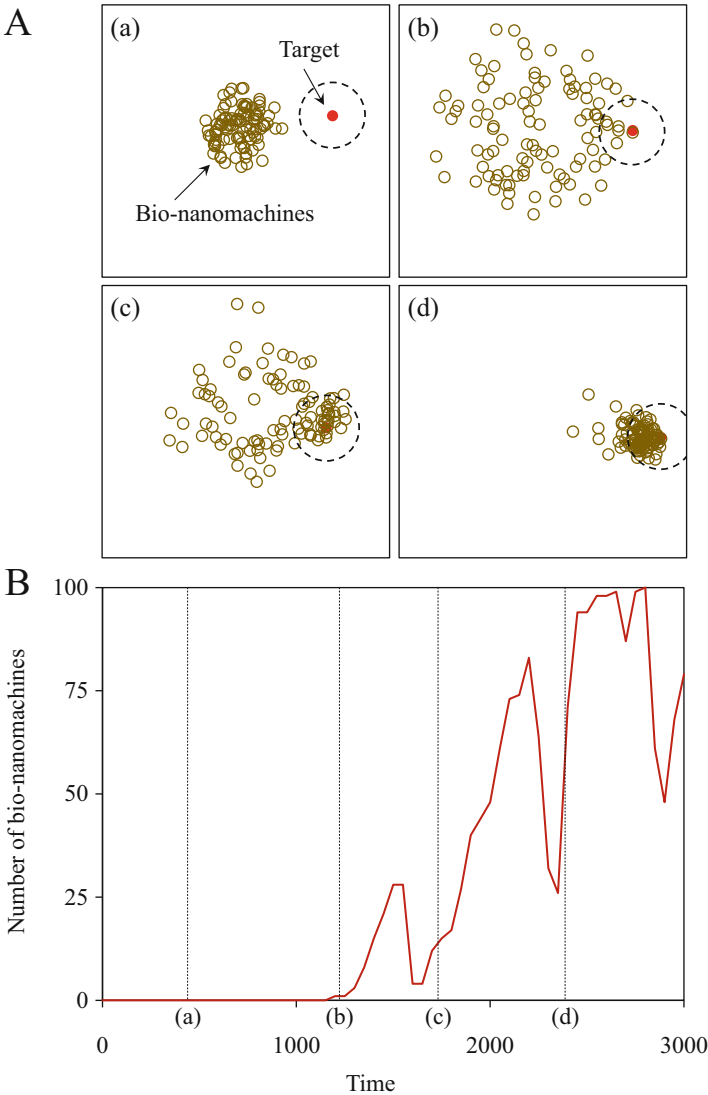
The mechanisms for target detection and tracking proposed in Okaie et al. (2016) use two types of molecule: repellents and attractants. In search of a target, bio-nanomachines release repellents to quickly spread in the environment; the released repellents form the concentration gradient in the environment, bio-nanomachines move toward lower concentrations of repellents, and therefore they move away from each other to help their search process. Upon detecting a

target, they release attractants to recruit other bio-nanomachines in the environment toward the target location; the released attractants also form the concentration gradient in the environment, and bio-nanomachines move toward higher concentrations of attractants, namely, toward the target location.

Figure 3A illustrates target detection and tracking processes. Here (a) a group of bio-nanomachines is placed in an environment where a single target exists; (b) the group of bio-nanomachines first spreads in the environment using repellents, and as a result one of the

Applications of Molecular Communication Systems, Fig. 3

Target detection and tracking processes (Okaie et al. 2016). (A) A group of 100 bio-nanomachines is placed in an environment containing a single moving target. The bio-nanomachines communicate to distribute in the environment and aggregates around the target location. (B) A performance measure is given by the number of bio-nanomachines in the proximity of the target



bio-nanomachines detects the target; (c) this bio-nanomachine, upon detecting the target, starts releasing attractants, and nearby bio-nanomachines are thus attracted to the location of the bio-nanomachine (i.e., near the target); and (d) as the target moves, the group of bio-nanomachines uses repellents and attractants in the same manner to move and track the target. Figure 3B shows how the number of bio-nanomachines in the proximity of the target (namely, the number of bio-nanomachines within the circle in Fig. 3B) changes with time. The up-and-down behavior of the graph indicates target detection and tracking processes that bio-nanomachines perform.

Future Directions

The current molecular communication research focuses more on theoretical work than experimental one; wet laboratory experiments form an importance piece of future work. Wet laboratory experiments help identify practical issues and gain insight into biologically implementable designs of molecular communication systems. An objective would be to show the collective behavior of bio-nanomachines for a specific application such as target detection and tracking.

Wet laboratory experiments are however challenging for communication engineers, since experimental facilities are often not available to them. To solve this problem, tabletop molecular communication platforms are developed in (Farsad et al. 2013). Tabletop molecular communication platforms are built using macroscale devices such as electronic sprays and sensors, yet communication is performed by propagating molecules between the macroscale devices. Tabletop MC platforms help communication engineers gain experience in molecular communication and develop realistic models to analyze the unique features and characteristics of molecular communication systems.

Nonetheless, theoretical work remains important in future work. Biologically realistic modelling and computer simulations will help

reduce the cost and time required to carry out wet laboratory experiments. This will also help us understand the detailed dynamics of molecular communication, such as spatiotemporal concentrations of molecules, which is difficult to observe in wet laboratory experiments. Future work should therefore use an integrated approach of experimental and theoretical investigations in order to design and develop practical applications of molecular communication systems.

Cross-References

- [Connectivity via Molecular Signaling](#)
- [Drug Delivery via Nanomachines](#)
- [Modeling Approaches for Simulating Molecular Communications](#)
- [Molecular Bit Detection](#)
- [Molecular Event Detection](#)
- [Modulation in Molecular Signaling](#)
- [Nanonetworks](#)
- [Neuronal Communication Channels](#)
- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)
- [Reaction-Diffusion Channels](#)

References

- Abdi A, Einolghozati A, Fekri F (2017) Quantization in molecular signal sensing via biological agents. *IEEE Trans Mol Biol Multi-Scale Commun* 3(2): 106–117
- Akyildiz IF, Jornet JM (2010) The Internet of nano-things. *IEEE Wirel Commun* 17(6):58–63
- Akyildiz IF, Pierobon M, Balasubramaniam S, Koucheryavy Y (2015) The Internet of bio-nano things. *IEEE Commun Mag* 53:32–40
- Bush SF, Paluh JL, Piro G, Rao V, Prasad RV, Eckford A (2015) Defining communication at the bottom. *IEEE Trans Mol Biol Multi-Scale Commun* 1(1):90–96
- Enomoto A, Moore M, Nakano T, Egashira R, Suda T, Kayasuga A, Kojima H, Sakakibara H, Oiwa K (2006) A molecular communication system using a network of cytoskeletal filaments. In: *Proceedings of 2006 NSTI nanotechnology conference and trade show*, vol 1, pp 725–728
- Farsad N, Eckford AW, Hiyama S, Moritani Y (2012) On-chip molecular communication: analysis and design. *IEEE Trans Nanobiosci* 11(3):304–314

- Farsad N, Guo W, Eckford AW (2013) Tabletop molecular communication: text messages through chemical signals. *PLOS ONE* 8(12):e82935
- Felicetti L, Femminella M, Reali G, Nakano T, Vasilakos AV (2014) TCP-like molecular communications. *IEEE J Sel Areas Commun (JSAC)* 32(12):2354–2367
- Hiyama S, Inoue T, Shima T, Moritani Y, Suda T, Sutoh K (2008a) Autonomous loading, transport, and unloading of specified cargoes by using DNA hybridization and biological motor-based motility. *Small* 4(4):410–415
- Hiyama S, Moritani Y, Suda T (2008b) Molecular transport system in molecular communication. *NTT DOCOMO Tech J* 10(3):49–53
- Kadloor S, Adve RS, Eckford AW (2012) Molecular communication using Brownian motion with drift. *IEEE Trans Nanobiosci* 11(2):89–99
- Kiourti A, Psathas KA, Nikita KS (2014) Implantable and ingestible medical devices with wireless telemetry functionalities: a review of current status and challenges. *Bioelectromagnetics* 35(1):1–15
- Lio P, Balasubramaniam S (2012) Opportunistic routing through conjugation in bacteria communication nanonetwork. *Nano Commun Netw* 3(1):36–45
- Mahfuz MU, Makrakis D, Mouftah HT (2010) On the characterization of binary concentration-encoded molecular communication in nanonetworks. *Nano Commun Netw* 1(4):289–300
- Moritani Y, Nomura S-iM, Morita I, Akiyoshi K (2010) Direct integration of cell-free-synthesized connexin-43 into liposomes and hemichannel formation. *FEBS J* 277:3343–3352
- Nakano T (2017) Molecular communication: a 10 year retrospective. *IEEE Trans Mole Biol Multi-Scale Commun* 3(2):71–78
- Nakano T, Koujin T, Suda T, Hiraoka Y, Haraguchi T (2009) A locally induced increase in intracellular Ca^{2+} propagates cell-to-cell in the presence of plasma membrane ATPase inhibitors in non-excitable cells. *FEBS Lett* 583(22):3593–3599
- Nakano T, Moore M, Wei F, Vasilakos AV, Shuai JW (2012) Molecular communication and networking: opportunities and challenges. *IEEE Trans NanoBiosci* 11(2):135–148
- Nakano T, Kobayashi S, Suda T, Okaie Y, Hiraoka Y, Haraguchi T (2014) Externally controllable molecular communication. *IEEE J Sel Areas Commun (JSAC)* 32(12):2417–2431
- Okaie Y, Nakano T, Hara T, Nishio S (2016) Target detection and tracking by bionanosensor networks. *Springer-Briefs in computer science*. Springer, Singapore
- Pierobon M, Akyildiz IF (2013) Capacity of a diffusion-based molecular communication system with channel memory and molecular noise. *IEEE Trans Inf Theory* 59(2):942–954
- Rogers U, shung Koh M (2016) Parallel molecular distributed detection with brownian motion. *IEEE Trans NanoBiosci* 15(8):871–880
- Tavakkoli N, Azmi P, Mokari N (2017) Performance evaluation and optimal detection of relay-assisted diffusion-based molecular communication with drift. *IEEE Trans NanoBiosci* 16(1):34–42
- Wei G, Bogdan P, Marculescu R (2013) Bumpy rides: modeling the dynamics of chemotactic interacting bacteria. *IEEE J Sel Areas Commun (JSAC)* 31(12):879–890

Architecture and Data Management for Smart Community Information Platform

Hiroaki Nishi

Department of System Design Engineering, Keio University, Yokohama, Japan

Synonyms

Data Privacy; Edge Computing; Fog Computing; Information Platform; Internet of Things (IoT); Smart City; Smart Community; Smart Community Services; Smart Infrastructure; Social Data Management; Vender and Consumer Relationship Management

Definition

Smart Infrastructure: A new style of infrastructure composed of conventional infrastructure and information and communication technologies (ICTs). Smart infrastructure refers to an efficient and highly functional infrastructure accomplished by ICT-based monitoring, control, and management.

Smart Community: An integration of smart infrastructures implemented in a certain region.

Introduction

A smart community is an integration of smart infrastructures implemented in a certain region. Each smart infrastructure may provide dedicated merits. However, the essential merit of a smart

community is community big data, which covers all infrastructures. To pursue the effective use of the community, big data is the key to the success of smart community projects. Therefore, data infrastructure is a necessity for a smart community. The information platform considering the safe data exchange and privacy preservation is described as a case study in Saitama City, Japan.

Smart Community

A smart city is an integration of smart infrastructures that are composed of conventional infrastructure and information and communication technologies (ICTs). Smart infrastructure refers to an efficient and highly functional infrastructure accomplished by ICT-based monitoring, control, and management (Nishi 2018). For example, a smart grid is composed of the interaction between the power grid and ICTs, and it accomplishes high-functioning grid operation and effective electricity usage. Smart transportation achieves automated drive and fare systems in logistics, cars, and road systems, which is also referred to as an intelligent transportation system (ITS). Smart agriculture improves the product value by controlling the growth of crops and the total efficiency of farm work. A smart government provides administrative services using the Internet to improve usability and operational efficiency. A smart city is an implementation of several smart infrastructures in a city. Similarly, a smart town and a smart island are implemented in a town and on an island, respectively. A smart community refers to a similar concept and is unrelated to the target region. Moreover, it is meaningless when these smart infrastructures are independently implemented in the target region. A smart community provides new services by integrating and linking information of different smart infrastructures. An example of this information integration is the efficient charge/discharge management for electric vehicles, combining data from an electric vehicle and the electric power system with traffic information. The intensive introduction of smart

infrastructures and strong information linkage enable the provision of more advanced services to communities. The penetration of Internet of Things (IoT) has created a large amount of data, and the success of the smart community project depends on the effective use of the data. The data management for a smart community is indispensable to provide attractive smart community services.

A smart community is described in the Smart Communities Guidebook (1997) by the State University of San Diego as “a geographical area ranging in size from neighborhood to a multi-county region whose residents, organizations, and governing institutions are using information technology to transform their region in significant ways. Co-operation among government, industry, educators, and the citizenry, instead of individual groups acting in isolation, is preferred. The technological enhancements undertaken as part of this effort should result in fundamental, rather than incremental, changes.” The ICT created a paradigm shift in infrastructures, and a significant amount of data processing and network transactions was generated. The data processing is primarily achieved in cloud services. However, the location of cloud causes obstacles to some services. The communication latency may cause a serious problem to hard real-time control applications. Open Fog (OpenFog Consortium 2018; Bonomi et al. 2012; Yi et al. 2015) proposed the placement of services closer to the terminal devices than the cloud for improving the efficiency of providing the services. Therefore, Fog would improve the service latency and improve the service distribution in the networks. Edge computing is used as a similar technology to Open Fog. Edge computing provides a computing environment at the edge of the Internet. This also means that the locations of the computing resources are closer to the IoT nodes than to the cloud. Data freshness is indispensable for some services, such as the ancillary service of power grids, and the automated car drive service. The study regarding Fog and Edge computing has become active as the discussions of new smart community services become popular. For data management,

FIWARE (<https://www.fiware.org>) (Ferreira et al. 2017) provides several types of data management API for smart community services. FIWARE provides various types of APIs to manage the smart community data; this shows the importance of data management, especially the multiple perspectives for the data management of smart city services. Herein, the smart community information platform from the perspective of its data management is focused and described.

Smart Community Data Specifications and Requirements

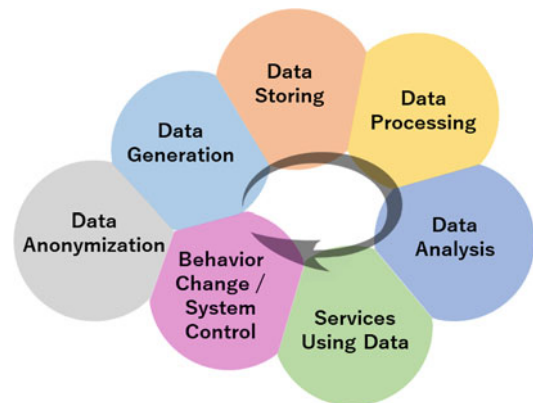
The requirements of smart city services are diverse. When providing smart city services, data specifications are required to provide the service. Several points for data specifications and requirements are used and described as follows.

Immediate data handling: Regarding time constraints, for example, ancillary services in a smart grid do not legally allow a delay of over 1 Hz. The IEC 61000-4 standard determines that the control delay has to be smaller than 10-ms intervals. Moreover, the IEC 61850-9-2 standard determines the tolerance of the delay to be less than 1 μ s. Thus, it is almost impossible to provide ancillary services as cloud services. However, it is necessary to accommodate new services that handle hard real-time system requirements, such as haptic communication for telemedicine and tactile sensing, which are sensitive to delay because they are required to transmit action-reaction force feedbacks to convey tactile sensations. This application also permits a 10-ms delay for maintaining stabilization in the control of its feedback control system (Anderson and Spong 1989). Cloud services cannot meet this need owing to their comparatively large communication delay.

Flexible data handling: Many protocols are proposed for communicating with IoT nodes. When focusing only on smart community protocols, especially the application layer, several examples can be provided: IEEE1888, IEEE1451, HTTP, MQTT, SEP2.0, Bluetooth (such as health thermometer profile, Bluetooth

Low Energy (BLE)), etc. In the application layer protocols, the primary purpose is to define the data formats. Therefore, server applications are required to support different protocols for receiving data from IoT devices, which differ in the types of application layer protocol. Moreover, new protocols and new data expression rules are successively developed to support the emerging smart community services. However, it is better to continually use conventional IoT devices that do not support new protocols because the replacement of all old IoT devices is costly. Moreover, it is typical that IoT terminals, which have similar sensors and functions, are installed redundantly at the same place owing to different protocols required to handle the data, and this poses a significant problem in the implementation of smart community services. It is important to design a smart community information infrastructure to address this problem.

Data cycle for improving QoL and QoS: For providing smart community services, it is essential to generate, store, process, and analyze the data, as well as foster the data to useful services, use the data through the services, change someone's behavior or control something using the data, and anonymize the data for its secondary use. Finally, the result of the behavior or control is measured by the sensors, and it creates the data again. This data cycle is shown in Fig. 1. This cycle enables the human behavior and machine



Architecture and Data Management for Smart Community Information Platform, Fig. 1 Smart community data cycle

control to be improved. Therefore, it improves the quality of life (QoL), such as the improvement in comfort or wellbeing, and the quality of service (QoS), such as the system efficiency or functionality. Thus, the smart city data infrastructure has to maintain the effective data cycle. The similar cycle is also discussed in ambient intelligence from a different perspective.

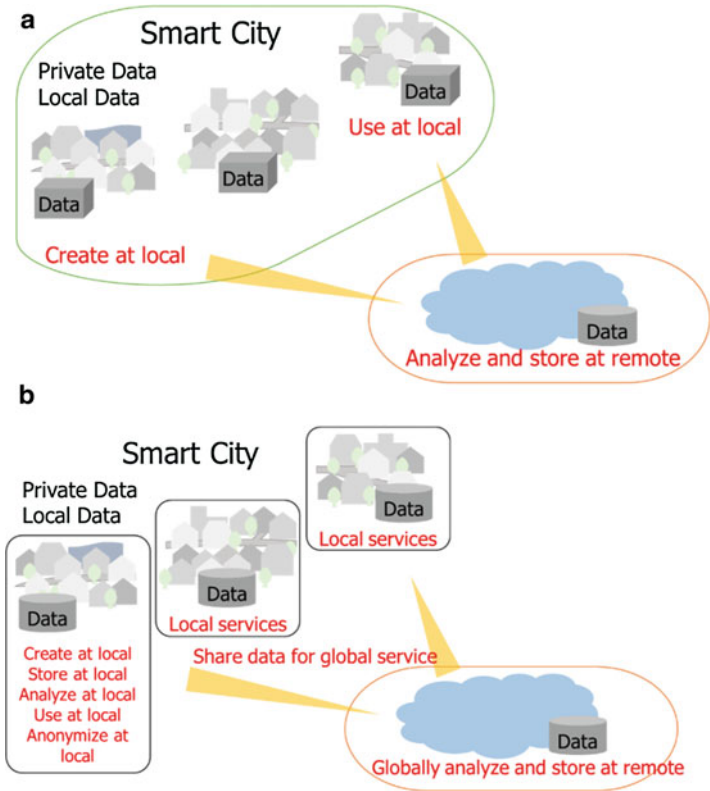
Data encapsulation: In the series of data processes, most parts of the process are achieved in the cloud, as shown in Fig. 2a. The use of cloud has the merit of low management cost. However, to use the cloud means to use a centralized system. Therefore, cloud services may cause a single point of failure and suffer from an advanced persistent threat. What types of countermeasures can be provided for it? It is strategically correct to prohibit gatherings and distribute people when the danger of terrorism is predicted. However, despite the repeated incidents of cyber terrorism, private data continue to be centrally managed in the cloud. It is also natural for people to

resist the revelation of personal information by moving these data from the cloud to their residential area. However, the service providers only consider their points of view and would force users to observe their service provision policy: “personal information must be managed in the cloud for providing services.” This situation is not desirable in the provision of smart community services. Private data should be encapsulated locally, and this encapsulation is effective from the perspective of protecting several types of attacks or threats, as shown in Fig. 2b. The cloud should provide a global service using abstract and unified information.

Data hierarchy: IoT devices send measured data to the cloud. On their way to the cloud, the data may pass a gateway at a smart house or smart building, or switches and routers at the internet providers, and finally, it will arrive at the service application programs in the cloud. As explained in data encapsulation, the data should be processed at an appropriate location when required

Architecture and Data Management for Smart Community Information Platform, Fig. 2 (a)

Conventional smart city data services. (b) Encapsulated smart city data services



Target Area	Narrow	Building <100 m	Town <10 km	City <100 km	Worldwide	Wide
Allowable Processing Delay	Short	Power System Stabilization Auto drive Collision Avoidance <10 ms		Facility Control <1 s Dynamic Pricing <30 min		Long
Calculation Cost Computing Platform	Small	Embedded Microcontroller	Server		Cluster Datacenter	Big
Anonymity	Weak	Plain	Weak Anonymization			Strong
			Strong Anonymization			
Amount of Data	Small	<KB	<MB		>GB	Big

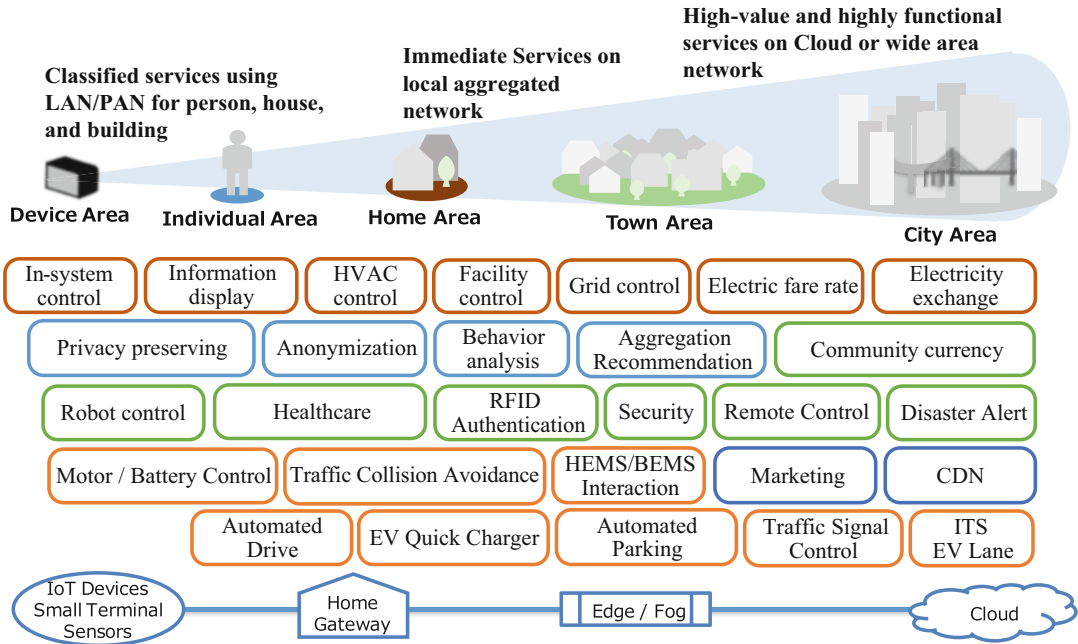


Architecture and Data Management for Smart Community Information Platform, Fig. 3 Smart community service hierarchy

by the service. As shown in Fig. 3, the switches and routers as well as the gateways can become computational resources referred as edge and fog computing. The selection of the appropriate computational resources requires appropriate indexes, and the typical indexes are shown in the figure. When the services are provided in a gateway, it focuses on narrow space service provisioning and short-range communications, e.g., an indoor environment, and the amount of calculation cost and computing platform can be small because the required data amount is small. Anonymity is a new and important index, as given in data encapsulation. The gateway is often connected with IoT devices and handles the raw data generated by the devices. Contrastingly, the services provided in the cloud can use anonymized data, and vice versa, as explained in data encapsulation. The Edge or Fog computing environment provides a balanced environment between the gateway and cloud. By using these indexes and the differentiated environment, data can be stored at an appropriate network location and processed at an appropriate network location. Moreover, the needs of data processing chains on the Internet

are increasing. This chain of data processing is also discussed as service function chaining in network function virtualization (NFV) (Trajkovska et al. 2017). This data chain is important in providing services using the data. Figure 4 shows the mapping of promising smart community services according to the indexes. For example, HVAC control is provided in a building and may use private information. Electricity exchange service is provided in the city area, and the price can be defined using generalized information. Additionally, the service applications have to select the best location.

IoT data security: The risk of IoT terminals being hacked has increased in the recent years. Hence, the maintenance of security levels must be ensured. However, most IoT terminals have low computing and power capacities. Therefore, it is difficult to facilitate better resistance to cyber-attacks using complicated protocols to add new functionalities. Moreover, introducing additional security software, or new protocols, is undesirable from the viewpoint of power consumption and system cost. Thus, it is necessary to design an information infrastructure to address these



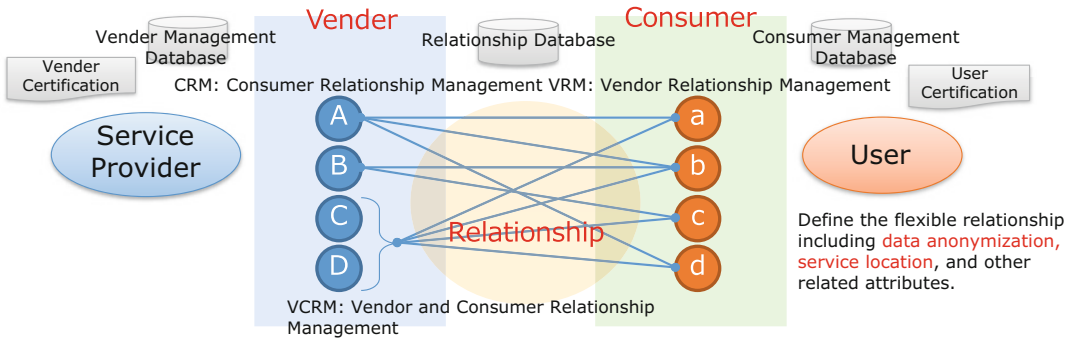
Architecture and Data Management for Smart Community Information Platform, Fig. 4 Mapping of smart community services considering service locations (Nishi 2018)

problems. Namely, the network system should provide new security services for IoT devices and locally stored data.

Private data handling: The protection of private information is important in smart community services because the data generated by residents for receiving smart community services could have privacy constraints. One of the protection methods is the aforementioned local data encapsulation. However, it is desirable to share the data for secondary use. This secondary use of data is the most promising service in the future smart community. Anonymization technology is key for the safe sharing of private data. If the data are perfectly anonymized and any type of private data is not extracted from the anonymized data, it is safe but is not useful for smart community services. The balance between difficulty in revealing private information and the usefulness in providing services is important. For achieving the balance at the highest level, the appropriate anonymization method must be designed and selected for each data service. Medical records, smart meters, locations,

and other data will have their respective suitable anonymization methods. The development of an anonymization method is indispensable. Moreover, watermarking technology for anonymized data is indispensable (Nakamura et al. 2017). In data service, the information of user rights must be managed, such as who generates the data, who uses the data, and what the purpose is. Watermarking technology can include the information in the anonymized data. The watermark in anonymized data can prevent data leakage by tracing the data leakage and protect the user rights holders.

Data ownership: Who possesses the data? Although this is a simple question, it is sometimes difficult to answer because the ownership and stakeholders of data are easy to entangle and handover. Moreover, the national speculations have worsened this situation. The General Data Protection Regulation (GDPR) was issued, and it provided a new stage of data management from the perspective of data privacy. In consideration of these situations, flexible data ownership management is required. Regarding the style of ser-



A

Architecture and Data Management for Smart Community Information Platform, Fig. 5 Vendor and consumer relationship management

vice provision, vendors have to manage the consumers; vendors draft contracts with consumers for providing services and provide reasonable prices for the services, in which the vendors aim to maximize both their benefit and the consumer satisfaction. In this contract, the vendors request the disclosure of private data. This is called consumer relationship management (CRM). Contrastingly, new trend requires an alternative relationship management between vendors and consumers such that the consumer can manage the vendors' services. This is called vendor relationship management (VRM). A user may think that a certain service is useful and worth providing the user's private data to the service vendor for a more efficient service. Meanwhile, the user may think that the service is enough to use as a trial and not worth providing the private information. In this case, the data management system will anonymize the data to preserve privacy. For attaining the flexible service relationship management between vendors and users, vendor and consumer relationship management (VCRM) is proposed (Niwa and Nishi 2017). Figure 5 shows the structure of VCRM. Both the vendor and consumer manage the database, which stores the information of the service provider or who the data user. A relationship is established when a contract of providing and receiving a service is engaged. The relationship is managed by its dedicated database, and the database has various information, such as the method and level of data anonymization, target area of service, location to be processed, service provider, service user, expi-

ration date, constraints for providing services, required computational power, required latency, and amount of data. VCRM manages these data and verifies the integrity with real use.

Smart Community Information Platform (SCIP)

The fusion of Fog/Edge, gateway, and cloud are appropriate as a desired future direction. However, their current shortfalls must be compensated. Therefore, an information mechanism called authorized stream contents analysis (ASCA) was proposed (Nishi 2018). ASCA is a mixed mechanism of software and hardware for supporting service provision. It supplies a method to process the information flowing through a communication network on intermediate communication devices. ASCA reconstructs the TCP stream on the device, decodes the SSL using the key shared with the cloud service, decodes Chunk and Gzip, executes string matching and the extraction function using regular expression rules, and subsequently sends the analyzed result of the stream to the dedicated service process. ASCA can modify the stream contents directly, under the constraint of the buffer size. When using 128 G of memory, it can analyze more than two million TCP streams simultaneously without the limitation of the TCP stream length by employing a dedicated context-switch technology. ASCA on a general enterprise server can provide 20

Gbps of processing performance using hardware accelerators, such as DPDK for the accelerating network throughput using userland zero-copy communication, HyperScan on the Intel Xeon Processor for accelerating the string matching function, and Intel QuickAssist Technology for accelerating the throughput of the encryption, decryption, compression, and decompression.

A service application using ASCA on a Fog/Edge terminal can process the network stream and provide dedicated services at an intermediate location on the route to the target in the cloud. This does not require any modification of the IoT terminals; the destination IP address of the data stream can maintain its original IP address of the target in the cloud. Moreover, ASCA provides the functionality to modify the stream contents, enabling the direct removal or anonymization of private information at the intermediate nodes. NEGI (Takagiwa and Nishi 2015) is an original library created for ASCA. It is designed with no acceleration and can therefore be utilized in any Linux-based environment, including an ARM-based embedded platform. The Intel Xeon platform can achieve 1 Gbps of streaming. DooR is a hardware accelerated library of the ASCA, and it achieves a 20-Gbps stream process. The proposed smart community information platform (SCIP) uses NEGI or DooR as basic libraries for stream processing. On these ASCA basic libraries, the user and service provider can design their services as a Docker container (Miura et al. 2017). Docker is a virtualized environment for executing application programs, and it reduces the cost of launching and terminating application software, compared with the conventional virtual machine environment. Because Docker provides a virtualized environment isolated from the host machine, the zero-copy architecture including the DPDK is not available for the application software as a Docker container. Therefore, Docker's shared memory option is used to communicate with a DPDK-supported NIC via shared memory to cope with this problem.

ASCA can handle the aforementioned smart community data specifications and requirements. ASCA provides a transparent environment as

network nodes. Because ASCA can monitor or modify a network stream at any point, it enables immediate data handling using zero-copy communication and hardware accelerators. It can also offer flexible data handling because it can provide services at the intermediate nodes on the Internet and change a communication protocol transparently. When ASCA is used as a basic library in the smart community data platform, it can be the key device for rotating the data cycle for improving the QoL and QoS. All the given functions in Fig. 1 can be achieved or measured using service applications with ASCA. Transparent data encapsulation can be achieved by designing the data firewall of the data anonymization application on the ASCA. Moreover, the hierarchical design of the ASCA-based encapsulation enables the data hierarchy. The firewall function and antivirus function on ASCA-supported devices can provide IoT data security. By implementing the appropriate applications on the ASCA, private data can be handled and the data can be anonymized. VCRM can be an application of ASCA. Therefore, either the user or the service provider can define and modify the relationship at any time. This feature adopts the opt-in and opt-out of data registration. The default value is given as the opt-in. The on-demand modification of the relationship achieves the opt-out.

Another important point is the mobility of the service applications. The initial allocation to an appropriate location and the subsequent execution of a service should ideally occur automatically, and be migrated as necessary, without the need for an explicit migration request. Executing this effectively requires managing several resources and activities, including memory, storage, CPU, communication, task allocation, and distribution, to manage the Docker containers with service applications. Thus, a mechanism for resource management that performs functions similar to the basic functions of an operating system is necessary. This resource management system is called the SCIP OS. Some basic functions of the SCIP OS resemble those of the orchestrator for Docker containers. However, the orchestrator only considers the management of the Docker containers, whereas the SCIP OS focuses on

the service applications more broadly from the perspectives of service feasibility, IoT feasibility, and future feasibility. Moreover, ASCA enabled wireless station can be a center of local data manager in the wireless environment. It gathers data from sensor nodes and offers a variety of benefits to local services in the reachable range of its wireless signals.

System Demonstration

The data anonymization and watermark insertion application of UDCMi, a smart town project in Misono Town, Saitama City, Japan, was demonstrated at the Global City Team Challenge EXPO (<https://pages.nist.gov/GCTC/>). This application uses the smart metering of smart houses in UDCMi. As a smart community service, a lifestyle recommendation service for eco-life is provided as a nudge service. In this demonstration scenario, the smart meter sends the data to the cloud, and the nudge report was automatically generated using machine-learning technology. En route to the cloud, the data are also captured, anonymized, and watermarked at the gateway by ASCA. The extension of the function of IoT devices was proven, and the effectiveness of the nudge report using anonymized data was compared with the other report using raw data.

Conclusion

Data-oriented smart community services are indispensable for the sustainable advancement of smart cities. The proposed smart community information platform is a system considering the given data specification and requirements. The complexity of data handling at the network infrastructure will be increased according to the development of a society structure. However, an information platform maximizing the flexible data management can solve various regional problems, including urbanization problems. The improvement of QoL and QoS, i.e., the ultimate goal of a smart community, is not achieved

by a single metric but by an interoperable data approach including the residents' behavior change and machinery control that enables hopeful societies to be established. Further developments of the architecture and data management of the smart community information platform are expected.

Cross-References

- [Advances in Distribution System Monitoring](#)
- [Cyber-physical Security in Smart Grid Communications](#)
- [Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future](#)
- [Privacy-Preserving Data Aggregation for Smart Grids](#)
- [Smart Metering, Specific Challenges, and Solution Approaches](#)
- [Ultra-reliable and Low-Latency Communications for the Smart Grid](#)

Acknowledgments This work was partially supported by MEXT/JSPS KAKENHI Grant (B) Numbers JP17H01739, and also by the Technology Foundation of the R&D project, "Design of Information and Communication Platform for Future Smart Community Services" by the Ministry of Internal Affairs and Communications of Japan.

References

- Anderson RJ, Spong MW (1989) Bilateral control of teleoperators with time delay. *IEEE Trans Autom Control* 34(5):494–501. <https://doi.org/10.1109/ICSMC.1988.754257>
- Bonomi F, Milito R, Zhu J, Addepalli S (2012) Fog computing and its role in the internet of things. In: *Proceedings of the first edition of the MCC workshop on mobile cloud computing*. ACM, New York, pp 13–16. <https://doi.org/10.1145/2342509.2342513>
- Ferreira D, Corista P, Gão J, Ghimire S, Sarraipa J, Jardim-Gonçalves R (2017) Towards smart agriculture using FIWARE enablers. In: *International conference on engineering, technology and innovation (ICE/ITMC)*, Madeira Island, pp 1544–1551. <https://doi.org/10.1109/ICE.2017.8280066>
- Miura T, Wijekoon JL, Prageeth S, Nishi H (2017) Novel infrastructure with common API using docker for scaling the degree of platforms for smart community services. In: *International conference on industrial informatics*, Emden, pp 474–479. <https://doi.org/10.1109/INDIN.2017.8104818>

- Nakamura Y, Nakatsuka Y, Nishi H (2017) Novel Method to Watermark Anonymized Data for Data Publishing. *IEICE Trans Inf Syst* E100-D(8):1671–1679
- Nishi H (2018) Information and communication platform for providing smart community services system implementation and use case in Saitama City In: *Proceedings of the 2018 IEEE international conference on industrial technology (ICIT)*, Lyon, pp 1375–1380. ISBN: 978-1-5386-4053-1/18/. <https://doi.org/10.1109/ICIT.2018.8352380>
- Niwa A, Nishi H (2017) An information platform for smart communities realizing data usage authentication and secure data sharing. In: *Fifth international symposium on computing and networking*, Aomori, pp 119–125. <https://doi.org/10.1109/CANDAR.2017.73>
- OpenFog Consortium. <https://www.openfogconsortium.org/#fog-computing>. Accessed 7 July 2018
- Smart Communities Guidebook: Building smart communities, how California's communities can thrive in the digital age. International Center for Communications, College of Professional Studies and Fine Arts, San Diego State University, San Diego, 1997
- Takagiwa K, Nishi H (2015) Local trend detection from network traffic using topic model and network router. In: *ICOMP' 15 – The 2015 international conference on internet computing and big data*, Las Vegas, pp 53–59
- Trajkovska I, Kourtis M-A, Sakkas C, Baudinot D, Silva J, Harsh P, Xylouris G, Bohnert TM, Koumaras H (2017) SDN-based service function chaining mechanism and service prototype implementation in NFV scenario. *J Comput Stand Interfaces Arche* 54(P4):247–265. Elsevier Science Publishers. <https://doi.org/10.1016/j.csi.2017.01.002>
- Yi S, Li C, Li Q (2015) A survey of fog computing: concepts, applications and issues. In: *Proceedings of the 2015 workshop on mobile big data*. ACM, New York, pp 37–42. <https://doi.org/10.1145/2757384.2757397>

Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks

Mugen Peng and Yaohua Sun
Key Laboratory of Universal Wireless
Communications (Ministry of Education),
Beijing University of Posts and
Telecommunications, Beijing, China

Synonyms

Multilayer wireless networks

Definitions

HetNets are seen as new network paradigm evolutions to the fifth-generation wireless systems, which can cost-efficiently improve system coverage, capacity, latency, and so on.

Historical Background

With the explosive increase in mobile data traffic, operators have to continuously improve network performance. As a promising solution, the Het-Net is a new technology that can cost-efficiently improve system coverage and capacity (Peng et al. 2015a). Communication nodes with high transmit power, such as MBSs, are deployed for large coverage and high capacity, while LPNs, such as SCAPs, help to supply service in the coverage of some hotspots. By deploying additional LPNs within the local-area range and bringing LPNs closer to end-UEs, HetNets can potentially improve spatial resource reuse and extend the coverage, thus allowing future cellular systems to achieve higher data rates while retaining the uninterrupted connectivity and seamless mobility of cellular networks.

The LPN is identified as one of the key components to increase the capacity of cellular networks in dense areas with high traffic demands. When traffic is clustered in hotspots, such LPNs can be combined with HPN to form a HetNet. HetNets have advantages of serving hotspot customers with high bit rates through deploying dense LPNs, providing ubiquitous coverage, and delivering the overall control signalings to all UEs through the powerful HPNs. Actuarially, too dense LPNs will incur severe interference, which restricts performance gains and commercial developments of HetNets. Therefore, it is critical to control interference through advanced signal processing techniques to fully unleash the potential gains of HetNets. The CoMP transmission and reception is presented as one of the most promising techniques in 4G systems. Unfortunately, CoMP has some disadvantages in real networks because its performance gain depends

heavily on the backhaul constraints and even degrades with increasing density of LPNs. Further, it was reported that the average SE performance gains from the uplink CoMP in downtown Dresden field trials were only about 20% with non-ideal backhaul and distributed cooperation processing located on the base station.

To overcome the SE performance degradations and decrease the energy consumption in dense HetNets, a new paradigm for improving both SE and EE through suppressing inter-tier interference and enhancing the cooperative processing capabilities is needed in the practical evolution of HetNets. Meanwhile, cloud computing technology has emerged as a promising solution for providing high energy efficiency together with gigabit data rates across software-defined wireless communication networks, in which the virtualization of communication hardware and software elements place stress on communication networks and protocols.

To achieve these goals, the C-RAN has been proposed as a combination of emerging technologies from both the wireless and the information technology industries by incorporating cloud computing into RANs (Peng et al. 2016a). C-RANs have come with their own challenges in the constrained fronthaul and centralized BBU pool. A prerequisite requirement for the centralized processing in the BBU pool is an interconnection fronthaul with high bandwidth and low latency. Unfortunately, the practical fronthaul in C-RANs is often capacity and time-delay constrained, which has a significant decrease on SE and EE gains.

Consequently, H-CRANs are proposed in Peng et al. (2014) as cost-effective potential solutions to alleviating inter-tier interference and improving cooperative processing gains in HetNets through combination with cloud computing. The motivation of H-CRANs is to enhance the capabilities of HPNs with massive MIMO and simplify LPNs through connecting to a “signal processing cloud” with high-speed optical fibers. As such, the baseband datapath processing and the radio resource control for LPNs are moved to the cloud server so as to take advantage of cloud computing capabilities.

Unfortunately, H-CRANs are still challenging in practice. First, since the location-based social applications become more and more popular, the traffic data over the fronthaul between RRHs and the centralized BBU pool surges a lot of redundant information, which worsens the fronthaul constraints. Besides, H-CRANs do not take full advantage of processing and storage capabilities in edge devices, such as RRHs and “smart” UEs, which is a promising approach to successfully alleviating the burden of the fronthaul and BBU pool. Moreover, operators need to deploy a huge number of fixed RRHs and HPNs in H-CRANs to meet the requirements of peak capacity, which makes a serious waste when the volume of delivery traffic is not sufficiently large.

To solve such challenges in H-CRANs, revolutionary approaches involving new RAN architectures and advanced technologies need to be explored. Fog computing is a term for an alternative to cloud computing that puts a substantial amount of storage, communication, control, configuration, measurement, and management at the edge of a network, rather than establishing channels for the centralized cloud storage and utilization (Bononi et al. 2012). Inspired by the advanced characteristics of fog computing and to alleviate the existing challenges of H-CRANs and take full advantages of local caching, the F-RAN architecture has been proposed in Peng et al. (2016b).

In this chapter, we are motivated to make an effort to offer a comprehensive survey on system architectures, key techniques, and future trends in heterogeneous cellular networks, including HetNets, H-CRANs, and F-RANs ‘in addition, all the abbreviations are listed (Table 1).

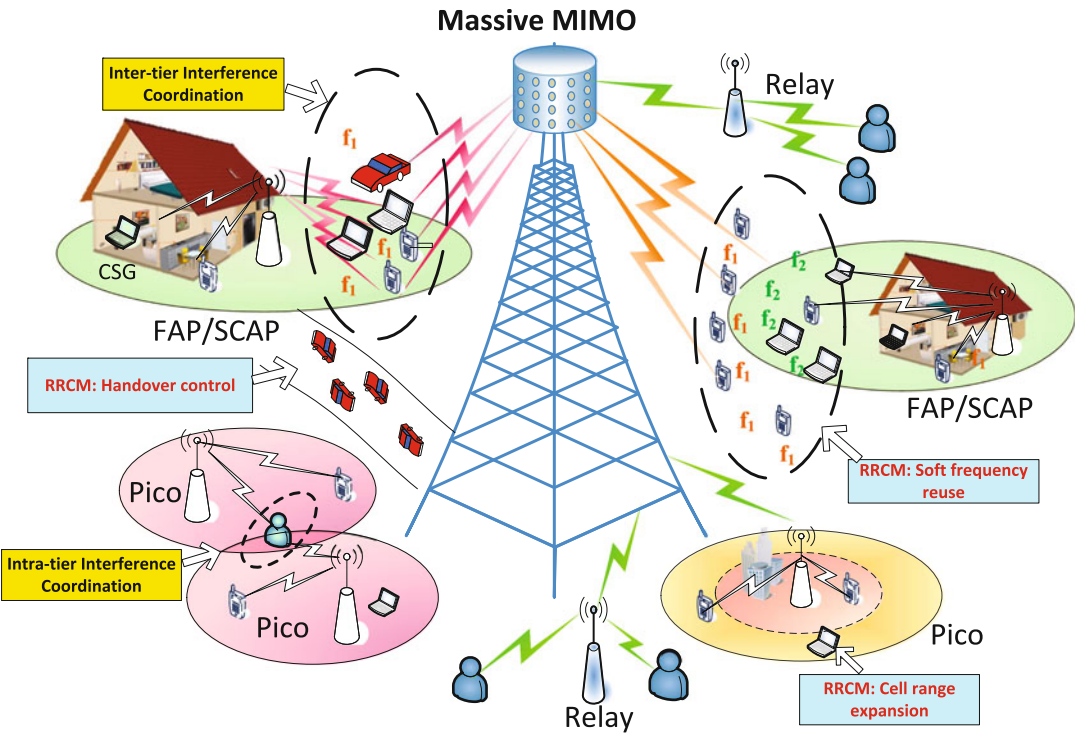
Foundations

The Architecture of Traditional HetNets

To provide a high-speed performance for cell-edge users, a hierarchical HetNet architecture Fig. 1 has been proposed for converging multiple LPNs and supporting unicast and multicast services simultaneously, where the coverage areas are divided into three layers: hierarchical

Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks, Table 1 Abbreviations in this chapter

3GPP	Third generation partnership project	JR	Joint reception
4G	Fourth generation	JT	Joint transmission
BBU	Baseband unit	LPN	Low power node
CNN	Convolutional neural network	LTE	Long-term evolution
CN	Core network	MBS	Macro base station
CoMP	Coordinated multi-point	MIMO	Multiple-input-multiple-output
CS/CB	Coordinated scheduling/coordinated beamforming	PHY layer	Physical layer
CRSP	Collaborative radio signal processing	QoE	Quality of experience
CRRM	Collaborative radio resource management	QoS	Quality of service
C-RAN	Cloud radio access network	RAN	Radio access network
DL	Downlink	RF	Radio frequency
DRL	Deep reinforcement learning	RNN	Recurrent neural network
EE	Energy efficiency	SCAP	Small cell access point
F-AP	Fog access point	SE	Spectral efficiency
F-RAN	Fog computing-based radio access network	SON	Self-organizing network
F-UE	Fog user equipment	UE	User equipment
H-CRAN	Heterogeneous cloud radio access network	UL	Uplink
HetNet	Heterogeneous network	WLAN	Wireless local area network
HPN	High power node		



Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks, Fig. 1 HetNet architecture for E-UTRAN systems

cooperative basic layer, homogeneous cooperative enhanced layer, and heterogeneous cooperative extended layer. For the hierarchical cooperative basic layer, UEs may be located near the BS; high-speed service is supported due to the utilizations of high-order modulation and coding schemes for the unicast services and hierarchical modulation schemes with unequal error protection space-time code for the multicast service. In the homogeneous cooperative enhanced layer, cooperative homogeneous diversity gain can be achieved, where UEs may be located near a cell boundary. For the heterogeneous cooperative extended layer, heterogeneous cooperative diversity gain guarantees the convergence and interworking of multiple RANs.

The Architecture of H-CRANs

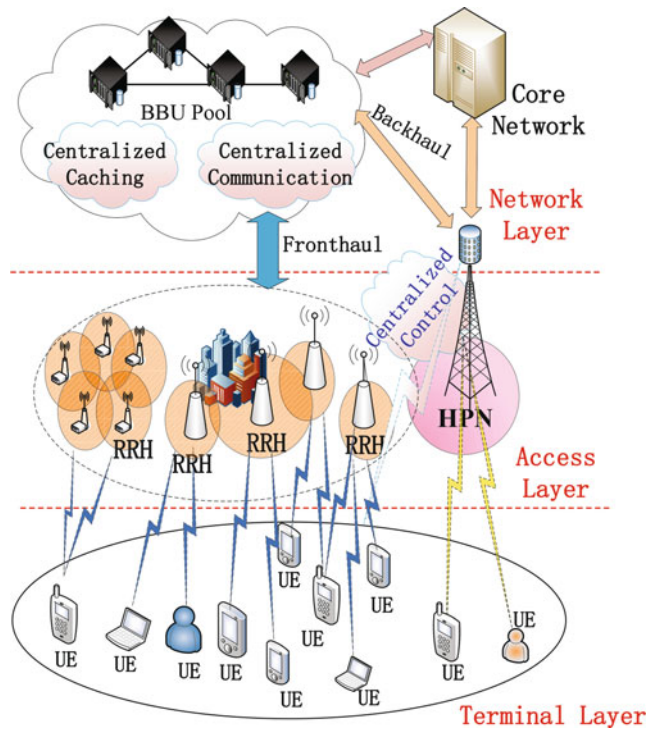
Similarly with the traditional C-RAN, As shown in Fig. 2, a huge number of RRHs with low energy consumptions in H-CRANs cooperate with each other in the centralized BBU pool to achieve high cooperative gains. Only the front RF and simple symbol processing functionalities are implemented in RRHs, while the other important

baseband physical processing and the procedures of the upper layers are executed jointly in the BBU pool. Sequently, only partial functionalities in the PHY layer are incorporated in RRHs, and the model with these partial functionalities is denoted as PHY_RF in Fig. 2.

However, different from C-RANs, the BBU pool in H-CRANs is interfaced to HPNs for mitigating the cross-tier interference between RRHs and HPNs through the centralized cloud computing-based cooperative processing techniques. Further, the data and control interfaces between the BBU pool and HPNs are added and denoted by S1 and X2, respectively, whose definitions are inherited from the standardization definitions of 3GPP. Since the voice service can be provided efficiently through the packet switch mode in 4G systems, the proposed H-CRAN can support both voice and data services simultaneously, and the voice service is preferred to be administrated by HPNs, while the high-data packet traffic is mainly served by RRHs.

Compared with the traditional C-RAN architecture, the proposed H-CRAN alleviates the fronthaul requirements with the participation

Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks, Fig. 2 System architecture of H-CRANs



of HPNs. Owing to the incorporation of HPNs, the control signalling and data symbols are decoupled in H-CRANs. All control signalling and system broadcasting information are delivered by HPNs to UEs, which simplifies the capacity and time-delay constraints of the fronthaul links between RRHs and the BBU pool, and can make RRHs active or sleep efficiently to save the energy consumption. Further, some burst traffic or instant messaging service with a small amount of data can be efficiently supported by HPNs. The adaptive signalling/control mechanism between connection-oriented and connectionless is supported in H-CRANs, which can achieve significant overhead savings in the radio connection/release by moving away from a pure connection-oriented mechanism. For RRHs, different transmission technologies in the PHY layer can be utilized to improve transmission bit rates, such as IEEE 802.11 ac/ad, millimeter wave, and even optical light. For HPNs, the massive MIMO is one potential approach to extend the coverage and enrich the capacity.

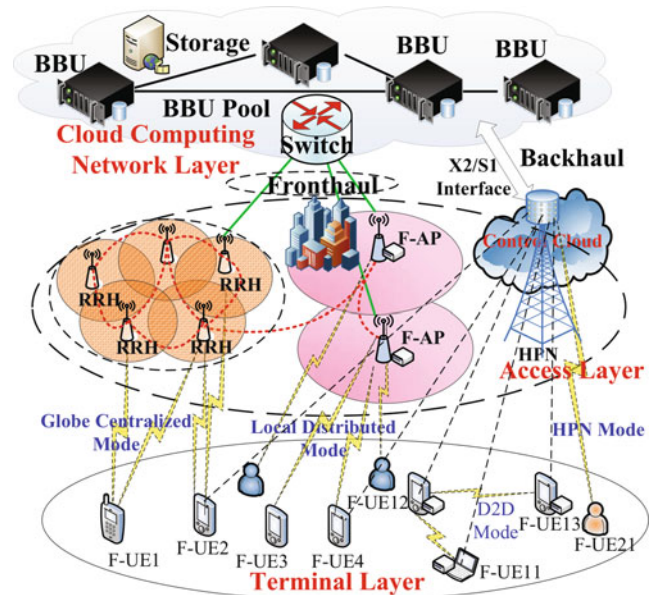
The Architecture of F-RANs

In the proposed F-RAN architecture shown in Fig. 3, the F-APs and F-UEs in the

terminal layer and network access layer constitute the fog computing layer, which are capable of conducting local caching, signal processing, and radio resource management. In the terminal layer, adjacent F-UEs can communicate with each other through the D2D mode or the F-UE-based relay mode. For example, F-UE13 and F-UE11 can communicate with each other with F-UE12 helping relay the signal, while F-UE11 and F-UE12 can operate in D2D mode to transmit data directly. In the network access layer consisting of F-APs and HPNs, HPNs are responsible for delivering control signalling to F-UEs, and F-APs are used to forward and process the received data. In addition, the data between F-APs and the BBU pool in the cloud computing layer is transmitted over fronthaul links, while HPNs are interfaced to the BBU pool with backhaul links for coordinations. The traditional X2/S1 interface for the backhaul link is backward compatible with that defined in 3GPP standards for LTE and LTE-Advanced systems.

Different from H-CRANs, a large number of CRSP and CRRM functions originally located in the cloud computing layer are shifted to F-APs and F-UEs, which alleviates the burden

Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks, Fig. 3 System model for implementing F-RANs



on the fronthaul and BBU pool. Meanwhile, the limited caching in F-APs and F-UEs can make some requests served locally. The benefit of this enhancement is significant, considering that the location-based social applications are becoming popular and much redundant information over the fronthaul links will worsen the fronthaul constraints. Furthermore, local caching, local signal processing, and local radio resource management can better satisfy the low latency requirement of some applications in 5G era.

Cooperative Spatial Processing

To boost the SE and EE performance of HetNets, the inter-tier interference should be properly handled. In the PHY layer, this interference can be mitigated by inter-tier CoMP transmission. The CoMP technique is aimed at enhancing the performance of cell-edge UEs. By the coordination between transmissions of adjacent cells, the CoMP can effectively suppress the adjacent cells' interferences when applied in UL or DL. Specifically, all access nodes can jointly decode transmitted signals from UEs in UL via JR while allocate transmission power to different UEs in a network MIMO manner with optimal precoding matrices, i.e., JT and CS/CB (Access 2010). It should be noted that the main difference between JT and CS/CB is that the data of each UE is transmitted by only one access node in CS/CB, while multiple cooperative nodes serve each UE simultaneously in JT. In Lee et al. (2012), it is demonstrated that JT and CS/CB can improve the performance of cell-edge UEs by 100% and 30%, respectively, compared to multiuser MIMO.

However, the backhaul transmission will incur latency for coordination in practice, and meanwhile the number of quantization bits for the exchanged information is limited, which will both degrade the performance of CoMP. The impacts of non-ideal backhaul are investigated in Xia et al. (2013), and it is concluded that the backhaul latency will affect the coverage and normalized throughput. Moreover, it is shown that the coverage and throughput achieved by CoMP zero-forcing beamforming decrease almost linearly with the average latency growing from zero. Even worse, when the latency exceeds

60% of the fading coherence time, the scheme does not have any benefit. Furthermore, the large number of cooperative cells does not essentially lead to better performance. On the contrary, it is shown that the number of coordinated cells should be kept fairly small in HetNets.

Radio Resource Management

To maximize the SE and EE performances of HetNets, the multidimensional radio resources should be optimized, including power allocation, subchannel allocation, and user association. However, because of the involvement of integer variables like subchannel allocation indicator and the existence of inter-tier and intra-tier interference, multidimensional radio resource allocation in HetNets is always non-convex, making the optimization problem hard to solve. To deal with the non-convexity, there are generally two kinds of ways. One is using heuristic algorithms which are suitable for finding the near-optimal solutions to NP-hard problems. The other one is to transform these non-convex problems into convex problems by different methods, and then the primal problems can be solved by settling the corresponding convex problems. In Peng et al. (2015b), the author focuses on maximizing EE by jointly allocating the resource block and transmit power for a heterogeneous cloud radio access network. Due to the non-convexity of the primal problem, the author tackles the primal problem by solving its dual problem aiming to minimize the Lagrange function of the primal problem which is always convex by definition. Moreover, since the primal problem satisfies the time-sharing condition, zero duality gap holds. Thus, the non-convex primal problem can be equivalently transformed into a convex optimization problem, i.e., its dual problem.

Furthermore, to achieve better global performance, it is very important to consider both the conventional physical layer performance metrics and the traffic delay in the upper layer. As a simple framework to deal with delay-aware RRM optimization problems, the Lyapunov optimization approach has been commonly adopted in HetNets. In Urgaonkar and Neely (2012), the Lyapunov optimization approach is utilized to

solve the problem of opportunistic cooperation in a two-tier HetNet. The access nodes would like to maximize their own throughput under average power constraints by optimizing access admission, cooperation decision, and power control. The obtained online control algorithm can stabilize the traffic queue without requiring any knowledge of the traffic arrival rates. While in Chen and Lau (2013), a two-stage queue-aware cross-layer resource management algorithm is proposed by minimizing drift-plus-utility. The cross-layer RRM consists of a queue-aware limited CSI feedback filtering optimization stage over a long timescale and an SINR-based optimal user scheduling stage over a short timescale. The utility is defined as the average feedback cost, and a large parameter V reduces the average CSI feedback at the cost of a larger average queue length.

Self-Organizing HetNets

Due to the ever-increasing number of radio resource management functionalities, it is impractical to manually adjust resource management parameters in a multidimensional parameter scenario. Faced with this issue, SON techniques have attracted a lot of attentions from both industry and academia. Generally speaking, the self-organizing capability of HetNets can be divided into three components: self-configuration, self-optimization, and self-healing. By using SON, the network configuration, installation, coverage, capacity, spectrum, and quality can be automatically managed and optimized taking into account target QoS performance, interference level, signal strength, traffic pattern, and so on.

To take full advantages of both centralized and distributed SON architectures, the author in Peng et al. (2013) proposes a hybrid SON architecture, and differences between traditional homogeneous networks and HetNets are discussed in terms of self-configuration like prediction-based radio resource configuration as well as self-optimization including mobility robustness optimization and energy saving. Moreover, it should be noted that time-varying traffic load makes the deployment of

completely self-organized HetNets nontrivial. Hence, the features of SON coordination in HetNets should be emphasized. Specifically, the graph-based decision framework proposed in Gelabert et al. (2014) can be adopted to enable efficient interaction and coordination of SON mechanisms, where the interaction and conflict relationship between multiple SON mechanisms are described by the metric event and action graph. Based on this proposed framework, the strategy-based SON coordination for HetNets can be easily implemented for a given event or a combination of events.

Key Applications

HetNets are the key to satisfying the diverse and stringent performance requirements in future wireless networks.

Future Directions

Access Slicing in HetNets

As a cost-efficient solution to support diverse use cases in the 5G era, network slicing enables the provision of networks in an as-a-service fashion. However, current studies mainly concentrate on CN-based slicing, and the impact of the characteristics of RANs is not well considered. To get a more effective network slicing solution, the author in Xiang et al. (2017) proposes an enhanced network slicing approach in F-RANs, termed as access slicing. The proposed access slicing architecture is featured with a centralized orchestration layer, a slice instance layer, and the information awareness capability to guarantee diverse QoS and QoE requirements. Although the above work makes a big step for network slicing, several challenges exist for its implementation. First, as stated in Xiang et al. (2017), the information awareness allows the access slice orchestration layer makes intelligent decisions on slice instance creation and resource management. Nevertheless, more specifications should be investigated like the types of information needed, the frequency for collecting information, and the way

of slice instance configuration based on processed information. Second, resource allocation between multiple slices should be investigated to fully utilize the resources of HetNets to meet diverse performance requirements. Such a problem can be very challenging, since it is basically a multi-objective optimization problem. Meanwhile, besides allocating radio resource, some slices also need caching resource and computation resource, which incurs an extra challenge.

HetNets Driven by Deep Learning

As an important technology for enabling artificial intelligence, deep learning has been successfully applied in many areas, including computer vision and speech recognition, and has drawn a lot of attentions of researchers in the wireless communication area. In Cao et al. (2017), a learning framework is proposed in which a CNN and an RNN are used to extract features in spatial domain and time domain from raw information collected by wireless networks, respectively, and this manner avoids identifying features manually. Taking the derived features as state input, DRL is adopted to control wireless networks intelligently, and the superiority of the proposal is verified by applying it to mobility management in WLAN. However, to apply deep learning in HetNets, more studies should be conducted. For example, more theoretical analysis should be done to provide guidelines on the architecture design of neural networks and the selection of hyper-parameters. Furthermore, the infrastructure of HetNets needs to be upgraded to support the implementation of deep learning algorithms, especially at the network edge to facilitate real-time network optimization and control.

Cross-References

- [Heterogeneous Small Cell Networks](#)
- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)
- [Resource Allocation](#)
- [Wireless Edge Caching](#)

References

- Access EUTR (2010) Further advancements for E-UTRA physical layer aspects. 3GPP technical specification TR 36:V2
- Bonomi F, Milito R, Zhu J, Addepalli S (2012) Fog computing and its role in the internet of things. In: Proceedings of the first edition of the MCC workshop on mobile cloud computing. ACM, Helsinki, Finland pp 13–16
- Cao G, Lu Z, Wen X, Lei T, Hu Z (2017) Aif: an artificial intelligence framework for smart wireless network management. *IEEE Commun Lett* 22:400–403
- Chen J, Lau VK (2013) Large deviation delay analysis of queue-aware multi-user MIMO systems with two-timescale mobile-driven feedback. *IEEE Trans Signal Process* 61(16):4067–4076
- Gelabert X, Sayrac B, Jemaa SB (2014) A heuristic coordination framework for self-optimizing mechanisms in LTE HetNets. *IEEE Trans Veh Technol* 63(3):1320–1334
- Lee D, Seo H, Clerckx B, Hardouin E, Mazzarese D, Nagata S, Sayana K (2012) Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges. *IEEE Commun Mag* 50(2):148–155
- Peng M, Liang D, Wei Y, Li J, Chen HH (2013) Self-configuration and self-optimization in LTE-advanced heterogeneous networks. *IEEE Commun Mag* 51(5):36–45
- Peng M, Li Y, Jiang J, Li J, Wang C (2014) Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies. *IEEE Wirel Commun* 21(6):126–135
- Peng M, Wang C, Li J, Xiang H, Lau V (2015a) Recent advances in underlay heterogeneous networks: interference control, resource allocation, and self-organization. *IEEE Commun Surv Tutor* 17(2):700–729
- Peng M, Zhang K, Jiang J, Wang J, Wang W (2015b) Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Trans Veh Technol* 64(11):5275–5287
- Peng M, Sun Y, Li X, Mao Z, Wang C (2016a) Recent advances in cloud radio access networks: system architectures, key techniques, and open issues. *IEEE Commun Surv Tutor* 18(3):2282–2308
- Peng M, Yan S, Zhang K, Wang C (2016b) Fog-computing-based radio access networks: issues and challenges. *IEEE Netw* 30(4):46–53
- Urgaonkar R, Neely MJ (2012) Opportunistic cooperation in cognitive femtocell networks. *IEEE J Sel Areas Commun* 30(3):607–616
- Xia P, Liu CH, Andrews JG (2013) Downlink coordinated multi-point with overhead modeling in heterogeneous cellular networks. *IEEE Trans Wirel Commun* 12(8):4025–4037
- Xiang H, Zhou W, Daneshmand M, Peng M (2017) Network slicing in fog radio access networks: issues and challenges. *IEEE Commun Mag* 55(12):110–116

Area Spectral Efficiency

► Area Spectral Efficiency of Ultradense Networks

Area Spectral Efficiency of Ultradense Networks

Guoqiang Mao
University of Technology Sydney, Sydney,
NSW, Australia

Synonyms

Area spectral efficiency; Ultradense networks

Definition

Area spectral efficiency refers to the data rate that can be achieved per unit bandwidth and in a unit area of the wireless network. It has the unit of b/s/Hz/m^2 or b/s/Hz/km^2 .

Historical Background

Network densification has been the main driver of wireless network capacity increase in the past and will play an even more crucial role in the development of the next generation mobile communication systems (5G). Network densification refers to the deployment of more base stations (BSs) and wireless access points per unit area and the associated technological advances to support such densification. There are three primary ways of increasing the wireless network capacity: (1) adding more spectrum, (2) enhancing spectral efficiency through advanced communication techniques, and (3) enhancing spatial reuse of frequency spectrum through network densification. The area spectral efficiency (ASE) is a major metric to measure the efficiency in the spatial

reuse of frequency spectrum. According to a study by Webb (Henderson 2007), among the 1-million-fold increase in wireless network capacity achieved in 50 years from 1950 to 2000, $15\times$ improvement was achieved from a wider spectrum, $5\times$ improvement from better Medium Access Control (MAC) and modulation schemes, $5\times$ improvement by designing better coding techniques, and an astounding $2700\times$ gain through network densification and reduced cell sizes. As we move to the next generation mobile communication systems and seek further capacity increases, network densification, manifested through heterogeneous and ultradense networks, will play an even more important role. First, the combined amount of spectrum available for licensed mobile broadband and unlicensed Wi-Fi is scarce. Spectrum approaching millimeter wavelengths is more abundant but is yet to be proven for cellular and Wi-Fi use due to difficulty in penetration and supporting significant mobility (Andrews et al. 2016a; Kutty and Sen 2016). Second, the scope for enhanced spectral efficiency may be limited as many current wireless systems are already running at a spectral efficiency close to the performance limit prescribed by the Shannon. Therefore, mobile operators have increasingly turned to network densification and small cell technology to meet the $1000\times$ capacity increase expected on 5G. Small cells are now deployed in massive numbers (Small Cell Forum 2016), which drives the paradigm shift into ultradense networks. Ultradense networks feature a very dense deployment of BSs and wireless access points.

Foundations

The ASE and wireless network capacity can be used interchangeably in the sense that the total capacity achieved by a network deployed in a given geographical area and using a specific amount of bandwidth is equal to the ASE times the size of the area and the amount of bandwidth. It is of crucial interest to investigate how the ASE

varies as more and more BSs are deployed and the network becomes denser and denser.

The SINR Invariance Principle for Low-to-Medium BS Density

Until recently, there was widely held belief that the ASE, or equivalently the network capacity, may increase indefinitely as the network densifies (Haenggi et al. 2009; Andrews et al. 2011). This view has been underpinned by the so-called SINR (signal-to-interference-plus-noise ratio) invariance principle, which is valid for a low-to-medium BS density. We use the following example to illustrate the SINR invariance principle. Consider a set of BSs deployed on an infinite plane, and number these BSs according to their distances to the origin such that the k -th nearest BS has a distance of l_k . Further assume that all BSs transmit at the same fixed power P_t and the wireless signal experiences standard power-law attenuation such that the received power at a distance d from the transmitter is $P_r(d) = P_t L d^{-\alpha}$, where L is a reference path loss at unit distance and α is the path loss exponent. For a “typical” user located at the origin and associated with its nearest BS, its SINR can be expressed as

$$\text{SINR} = \frac{P_t L l_1^{-\alpha}}{\sum_{k=2}^{\infty} P_t L l_k^{-\alpha} + \sigma^2}, \quad (1)$$

where σ^2 represents the noise power. Now consider scaling the distances between all BSs by a factor of t . The density of BSs will increase by a factor of $\frac{1}{t^2}$. Further assume that noise power is negligible, which is valid in an interference-limited regime where most wireless networks now operate in. The SINR then becomes

$$\begin{aligned} \text{SINR} &= \frac{P_t L (t l_1)^{-\alpha}}{\sum_{k=2}^{\infty} P_t L (t l_k)^{-\alpha} + \sigma^2} \\ &\simeq \frac{P_t L (t l_1)^{-\alpha}}{\sum_{k=2}^{\infty} P_t L (t l_k)^{-\alpha}} \\ &= \frac{l_1^{-\alpha}}{\sum_{k=2}^{\infty} (l_k)^{-\alpha}}, \end{aligned} \quad (2)$$

i.e., the SINR is invariant with the increase of BS density.

The SINR invariance principle implies that as the network densifies and the distances between transmitters and receivers reduce, the increase in interference will be counterbalanced by the increase in the desired signal. Consequently, the SINR will stay approximately the same. The principle suggests that other things being equal, the spectral efficiency, or equivalently the capacity, per BS cell is invariant as network densifies. Therefore, from the network perspective, the ASE, or equivalently the overall network capacity, will linearly increase with the number of cells per unit area; from the user perspective, as each user maintains the same SINR but shares its BS with an ever-smaller number of other users, each user can achieve approximately linear growth in its achievable data rate as BSs are added, until the limit of one user per cell is reached. The SINR invariance principle is not affected by the BS layout, transmit power, shadowing and fading distributions, and path loss exponent (Andrews et al. 2016b).

Area Spectral Efficiency for Ultradense BS Regime

Recent research suggested however that the SINR invariance principle may no longer apply when the BS density becomes very large and that there may exist a limit in network densification, beyond which further densification will not bring the expected linear increase in capacity and may even reduce the capacity (Ding et al. 2015, 2016a,b; Ge et al. 2016; Liu et al. 2016a,b; Andrews et al. 2016b; Zhang and Andrews 2015). Specifically, by incorporating those effects that have negligible impacts on the ASE when the BS density is small or moderate but become dominant factors determining the ASE when the BS density is very large, it was shown that the ASE may either exhibit a sublinear increase with the BS density, or reduce at certain region of the BS density, or even monotonically reduce to zero beyond a certain BS density threshold.

In Ding et al. (2015, 2016a,b), and Ge et al. (2016), researchers considered the impact of non-line-of-sight (NLoS) and line-of-sight

(LoS) transmissions on the ASE. NLoS and LoS transmissions are ubiquitous in wireless communications. Other things being equal, as the distance between a transmitter and a receiver increases, the probability that their direct LoS path gets blocked increases and the converse. NLoS transmissions will generally suffer much higher path loss than LoS transmissions. Specifically, by employing a 3GPP-endorsed LoS/NLoS probability model given below

$$\Pr^L(x) = \begin{cases} 1 - \frac{x}{d_1}, & 0 < x \leq d_1 \\ 0, & x > d_1 \end{cases}, \quad (3)$$

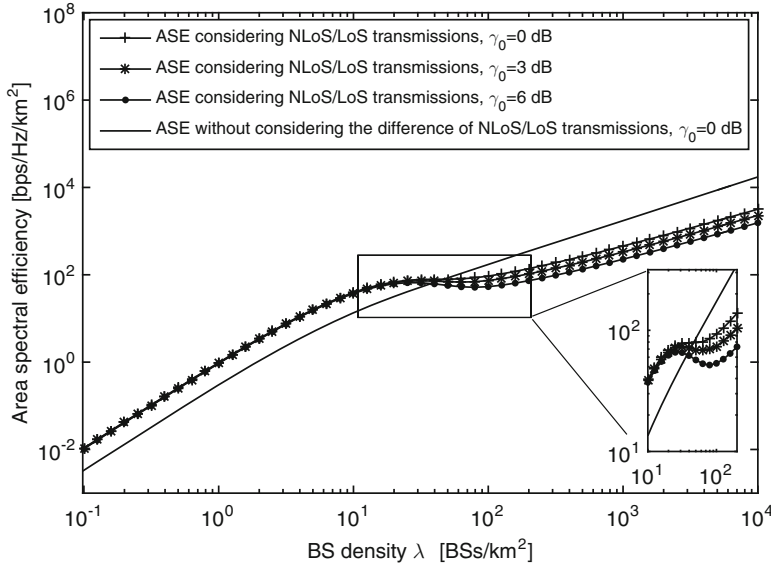
where $\Pr^L(x)$ is the probability that a transmitter and a receiver separated by a distance x experience LoS transmission, $1 - \Pr^L(x)$ is the probability that the same pair of transmitter and receiver experiences NLoS transmission, and d_1 is a parameter determining the decreasing slope of the linear function $\Pr^L(r)$. Ding et al. (2016a) investigated the variation of the ASE with the BS density, shown in Fig. 1. The ASE is determined using the following equation

$$ASE = \lambda \Pr(\text{SINR} \geq \gamma_0) \int_{\gamma_0}^{\infty} \log_2(1+x) f_{\text{SINR}}(x) dx \quad (4)$$

where λ is the BS density, γ_0 is the SINR threshold required for establishing a connection, $f_{\text{SINR}}(x)$ is the probability density function of the SINR, and the term $\log_2(1+x)$ comes from the Shannon capacity formula. In the literature, the ASE has also been calculated using the following formula $ASE = \lambda \Pr(\text{SINR} \geq \gamma_0) \log_2(1+\gamma_0)$, which reflects the fact that some wireless systems may not be able to explore the extra SINR above the SINR threshold γ_0 to boost the data rate.

Figure 1 shows that when LoS/NLoS transmissions are considered, the ASE exhibits not only quantitatively but also qualitatively different trends with the BS density, compared with the case that does not consider the distinction of LoS and NLoS transmissions. Specifically,

1. when the BS density is small, the ASE quickly increases with the BS density because the network is generally noise-limited and adding more BSs immensely benefits the ASE by reducing the transmission distance between the BS and the mobile user (MU). Most transmissions in this regime are NLoS transmissions; however the transmission between a MU and its desired BS has a higher probability of being LoS transmission, and as the BS density further increases, this probability increases to a non-negligible value. Therefore, in this region, the ASE considering LoS/NLoS transmissions is higher than that without considering LoS/NLoS transmissions, but their trend is the same;
2. when the network is dense enough, i.e., greater than 20 BS/km² in Fig. 1, the growth of the ASE with the BS density, when considering LoS/NLoS transmissions, becomes flat or even exhibits a decrease (for $\gamma_0 = 3$ dB and $\gamma_0 = 6$ dB), which is distinctly different from that predicted without considering LoS/NLoS transmissions. This can be explained by the so-called NLoS-LoS transition effect. Particularly, in this region, the transmissions between MSs and their desired BS are already dominated by LoS transmissions but in the beginning signals from interfering BSs, are mainly NLoS transmissions suffering higher loss. As BS density further increases and the distances between a MU and BSs, including both the desired BS and interfering BSs, further reduce, signals from interfering BSs start to transit from NLoS transmissions to LoS transmissions. Consequently, as the BS density increases, interference experiences a larger increase compared with the desired signal, causing the SINR to reduce sharply in this region. Therefore, the ASE remains largely flat with the increase in BS density or may even reduce;
3. when the network becomes very dense, i.e., greater than 100 BS/km² in Fig. 1, the dominant interfering BSs have completed their NLoS-LoS transitions. Both the desired signal and the dominant interference are now



Area Spectral Efficiency of Ultradense Networks,

Fig. 1 The variation of the ASE with the BS density at various SINR thresholds. Both the case considering the coexistence of NLoS and LoS transmissions and the case without considering the coexistence of NLoS and LoS transmissions are shown. The case without considering the coexistence is plotted using the result in Andrews et al. (2011) and assuming $\alpha = 3.75$, $P_t = 24$ dBm, $\sigma^2 =$

-95 dBm, and $L = 10^{-14.54}$. The case considering the coexistence of NLoS and LoS transmissions is obtained assuming $d_1 = 300$ m, $P_t = 24$ dBm, $\sigma^2 = -95$ dBm, $\alpha = 2.09$, and $L = 10^{-10.38}$ for LoS transmissions and $\alpha = 3.75$ and $L = 10^{-14.54}$ for NLoS transmissions. These parameters are chosen following 3GPP recommendations

LoS transmissions. However, non-dominant interfering BSs further away from the MS may still experience the NLoS-LoS transition effect. Therefore, in this region, the SINR may still reduce modestly with the increase in BS density. This modest decrease in the SINR combined with the increase in the BS density causes the ASE to increase but at a rate below that predicted without considering LoS/NLoS transmissions.

Open data at Ofcom, an independent regulator and competition authority for the UK (<http://sitefinder.ofcom.org.uk/search>), shows that in some dense urban regions in the UK, the BS density already exceeds 10 BS/km². As we move into the regime of ultradense networks, we are bound to pass through the aforementioned second region. Therefore, it is important to consider the effect of LoS and NLoS transmissions when analyzing the ASE of ultradense networks.

In 2016b, Liu et al. investigated the ASE from a different perspective by considering the transition from far-field to near-field radio propagation which may be triggered as the BS density increases to a very large value. Specifically, it is well known that wireless signal attenuates faster at larger distances. In a region close to the transmitter, the signal may suffer little attenuation, while this attenuation sharply increases as the distance from the transmitter becomes large. Based on some field measurements, they proposed the following multi-slope BPM model to better capture the signal attenuation

$$g_N^B(\{\alpha_n\}_{n=0}^{N-1}; x) = \eta_n (1 + x^{\alpha_n})^{-1},$$

$$R_n < x \leq R_{n+1}, \quad (5)$$

where x denotes the distance from the transmitter to the receiver, R_n separates the attenuation func-

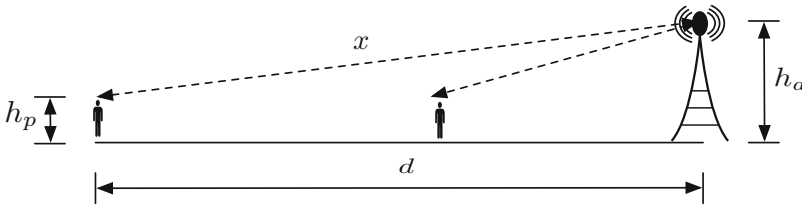
tion into several distinct regions, and α_n denotes the path loss exponent for $R_n < x \leq R_{n+1}$, $\eta_0 = 1$, and $\eta_n = \prod_{i=1}^n \frac{1+R_i^{\alpha_i}}{1+R_i^{\alpha_{i-1}}}$. As signals are attenuated faster at larger distances, $\alpha_n < \alpha_{n+1}$ holds. Using the model, they predicted that the ASE may monotonically decrease to zero when the BS density increases beyond a certain critical threshold. When the SINR threshold $\gamma_0 = 20$ dB, this critical BS density lies between 5000 – 25,000 BS/km² depending on the values of α_i ; when $\gamma_0 = 0$ dB, this critical BS density lies between $2 \times 10^5 - 3 \times 10^5$ BS/km². In a separate work (Zhang and Andrews 2015), Zhang and Andrews used a dual-slope (power-law) path loss function to model the different radio propagations in a region close to the transmitter and in a region far away:

$$l_2(\alpha_1, \alpha_2; x) = \begin{cases} x^{-\alpha_0} & x \leq R_c \\ \eta x^{-\alpha_0} & x > R_c \end{cases}, \quad (6)$$

where $\eta = R_c^{\alpha_1 - \alpha_0}$, $R_c > 0$ is the critical distance, and α_0 and α_1 are the near- and far-field path loss exponents with $0 \leq \alpha_0 \leq \alpha_1$. Using the model, they predicted that the ASE grows linearly with the BS density λ if $\alpha_0 > 2$, scales sublinearly with rate $\lambda^{2-\frac{2}{\alpha_0}}$ if $1 < \alpha_0 < 2$, and decays to zero if $\alpha_0 < 1$. It is worth noting that compared with extensive studies on far-field propagation models, surprisingly little is known about near-field radio propagation. Therefore, the models in (5) and (6) are plausible. However, a

common insight revealed in their work (Zhang and Andrews 2015; Liu et al. 2016b) is that the ASE exhibits quantitatively and qualitatively different trends with the BS density λ when the different propagation conditions in the near-field and in the far-field are considered.

In Ding and Perez (2016) and Gruber (2016), researchers considered the impact of some geometric conditions on the ASE. Specifically, in Ding and Perez (2016), Ding and David Lopez investigated the impact of antenna height on the ASE by considering a scenario where the BS antenna height is kept constant when its density increases. They showed that when the height of the BS relative to the MU is maintained at a constant value of 8.5 m, the ASE may peak when the BS density increases to ~ 15 BS/km² and then starts to monotonically decrease to zero as the BS density further increases. In comparison, when the impact of this height is ignored, the ASE may increase monotonically toward infinity. Their result can be intuitively explained using Fig. 2, which shows that while the impact of antenna height may be small at small BS density, its impact may play a dominant impact on the ASE in an ultradense network with a large BS density. In 2016, Gruber considered the geometric constraints limiting the BS deployment locations. Specifically, by considering that BSs can only be deployed along the two sides of the street, some distinctly different results on the ASE were obtained compared with that assuming BSs can be deployed randomly and anywhere in the street. It is worth noting that while some assumptions



Area Spectral Efficiency of Ultradense Networks, Fig. 2 An illustration of the different impacts of the antenna height as the distance between MU and BS changes. When the distance d between the MU and BS is large, the impact of the relative height between BS antenna, h_a , and MU, h_p , on the propagation distance x

is small and negligible. When the BS density increases to a very large value and hence the distance d reduces to a small value, the impact of $h_a - h_p$ on the propagation distance x is no longer negligible and may even be the dominant factor determining x

used in these studies are quite plausible, e.g., as the density of BS increases, one would expect that the BS size and height also reduce instead of being constant; the insight revealed in these studies holds: as we move into the ultradense regime, such geometric restrictions as physical dimensions of transmitters and receivers and deployment restrictions on BSs that previously have small or negligible impact on the ASE may now become dominant factors in the consideration.

In summary, distinct from conventional low-to-medium density networks, in ultradense networks, impacts of LoS/NLoS transmissions, the different radio propagation conditions in the near-field and far-field of BSs, and network geometric constraints may play an important role in determining the ASE.

Key Applications

The ASE is a fundamental performance metric of ultradense networks and more general wireless networks. The ASE determines the capacity that can be achieved by a wireless network and measures the efficiency in the spatial reuse of frequency spectrum. Knowledge of the ASE helps to guide the design and deployment of wireless networks.

Cross-References

- [Area Spectral Efficiency](#)
- [Cognitive Heterogeneous Networks](#)
- [Network Capacity](#)

References

Andrews JG, Baccelli F, Ganti RK (2011) A tractable approach to coverage and rate in cellular networks. *IEEE Trans Commun* 59(11):3122–3134

Andrews JG, Gupta AK, Dhillon HS (2016a) A primer on cellular network analysis using stochastic geometry. eprint arXiv:1604.03183

Andrews JG, Zhang X, Durgin GD, Gupta AK (2016b) Are we approaching the fundamental limits of wire-

less network densification? *IEEE Commun Mag* 54(10):184–190

Ding M, Lopez-Perez D, Mao G, Wang P, Lin Z (2015) Will the area spectral efficiency monotonically grow as small cells go dense? In: *IEEE GLOBECOM*, pp 1–7

Ding M, Perez DL (2016) Please lower small cell antenna heights in 5G. In: *IEEE Globecom*, pp 1–6

Ding M, Wang P, López-Pérez D, Mao G, Lin Z (2016a) Performance impact of LoS and NLoS transmissions in dense cellular networks. *IEEE Trans Wirel Commun* 15(3):2365–2380

Ding T, Ding M, Mao G, Lin Z, López-Pérez D (2016b) Uplink performance analysis of dense cellular networks with LoS and NLoS transmissions. In: *IEEE ICC*, pp 1–6

Ge X, Tu S, Mao G, Wang CX, Han T (2016) 5G ultra-dense cellular networks. *IEEE Wirel Commun* 23(1):72–79

Gruber M (2016) Scalability study of ultra-dense networks with access point placement restrictions. In: *IEEE ICC workshops*, pp 650–655

Haenggi M, Andrews JG, Baccelli F, Dousse O, Franceschetti M (2009) Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE J Sel Areas Commun* 27(7):1029–1046

Henderson W (2007) Webb report, ofcom.

Kutty S, Sen D (2016) Beamforming for millimeter wave communications: an inclusive survey. *IEEE Commun Surv Tutor* 18(2):949–973

Liu J, Sheng M, Liu L, Li J (2016a) Effect of densification on cellular network performance with bounded pathloss model. *IEEE Commun Lett* 21(2):346–349

Liu J, Sheng M, Liu L, Li J (2016b) How dense is ultra-dense for wireless networks: from far- to near-field communications. eprint arXiv:1606.04749

Small Cell Forum (2016) Small cell market status report, May 2016

Zhang X, Andrews JG (2015) Downlink cellular network analysis with multi-slope path loss models. *IEEE Trans Commun* 63(5):1881–1894

Artificial Intelligence

- [Machine Learning in Wireless Sensor Networks for the Internet of Things](#)

Artificial Jamming schemes

- [Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks](#)

Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks

Yi-Liang Liu

Department of Electronics Information Engineering, Harbin Institute of Technology, Harbin, China

Synonyms

Artificial jamming schemes; Physical layer security

Definition

Cellular communications and networks are particularly vulnerable to eavesdropping attacks due to the opened and broadcasting nature of wireless channels. Due to the rapid development of cellular networks and wireless business, the security issue has attracted a lot of attention. Physical layer security takes the advantages of channel randomness nature of transmission media to achieve communication confidentiality, which is the most important and interesting topic of privacy communication technologies. Artificial noise (AN) or jamming signals with MIMO technologies are supposed to implement physical layer security and enhance secrecy capacities in cellular networks, where AN signals along with confidential data confuse potential eavesdroppers via utilizing orthogonal spaces provided by transmit antennas. In these AN-based schemes, message streams were sent in a multiplexing mode via all eigen-subchannels (positive eigenvalue channels) at desired directions, and the AN signals can be transmitted to the null spaces (zero eigenvalue channels) of desired directions, so that they do not affect the desired user, while eavesdropper channels are degraded with a high probability. This chapter introduces a physical layer security review, covering the information theory and physical layer security schemes based on MIMO

technologies, and provides an overview on the state-of-the-art works on the AN-based scheme along with conclusions and future research directions.

Historical Background

The origin of physical layer security research can be traced back to Wyner's definition of an information-theoretic secrecy capacity (Wyner 1975), which is a maximum message transmission rate in confidential communications. Compared to conventional cryptography that works to ensure all involved entities to load proper and authenticated cryptographic information, physical layer security technologies perform confidentiality functions without considering about how those security protocols are executed. In other words, it does not require to implement any extra security schemes or algorithms on other layers above the physical layer. In addition, the physical layer has the same confidentiality level with the one-time pad, which ensures its security performance. The concept of physical layer security has become more popular with the help of MIMO technologies, because these emerging technologies can improve the secrecy capacity massively via utilizing extra orthogonal spaces provided by multiple antennas.

Information Theory

Wyner implied that the intrinsic and unpredictable elements of physical channels, such as noises, interferences, and wireless fading, play central roles in protecting secure messages (Wyner 1975) and claimed that there exists a randomized codebook maximizing transmission rates that can achieve perfect secrecy where any eavesdroppers cannot get any information. The maximizing secrecy rate is defined as a secrecy capacity C_s as

$$C_s = \max_{V \rightarrow X \rightarrow YZ} I(V; Y) - I(V; Z), \quad (1)$$

where $I(\cdot; \cdot)$ denotes mutual information between V and Y and V is an auxiliary input

variable with a joint auxiliary input distribution $p(v)p(x|v)$. Given a discrete memoryless channel $P_{YZ|X}$, wiretap codes achieve the secrecy capacity via maximization over the choices of the joint distribution $P_{YZ|X}$, such that the Markov chain $V \rightarrow X \rightarrow YZ$ holds, which has been expected to completely address the underlying problem of that traditional cryptographic algorithms are vulnerable to quantum attacks.

In the last two decades, researchers have developed a significant amount of mathematical theories, such as secrecy capacity characterization, wiretap coding designs, wireless fading managements, etc. And the advancements in cellular technologies have improved physical layer security significantly, by exploring spatial diversities and multiplexing gains with the help of multi-antennas technologies.

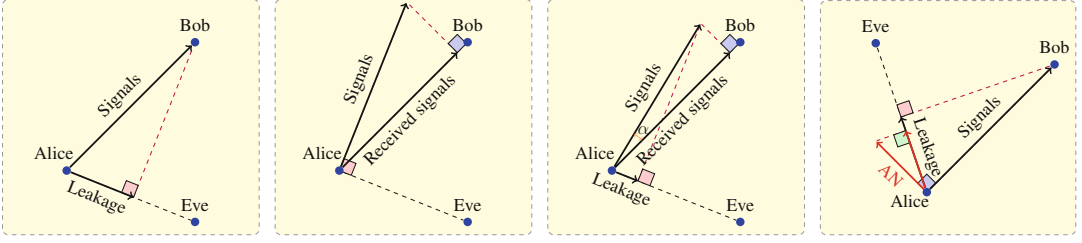
Physical Layer Security Schemes Based on MIMO Technologies

From the traditional communication viewpoints, improving the link quality of the main channel is one of the most feasible approaches to improve the secrecy capacities with existing installations of cellular networks, because most of the cellular wireless communication technologies can improve the channel spectrum utilization, and if the eavesdropping channel spectrum utilization is unchanged, these traditional wireless communication technologies will undoubtedly improve the secrecy capacities. However, this viewpoint has a great security risk, i.e., the secrecy capacities are small if the eavesdropper has a better channel quality than the desired users. Therefore, current research insists on another viewpoint that jointly uses more degree of freedom of main channels and artificial randomness to protect messages while reduces the quality of eavesdropping channels. Prospectively, the randomness of multipath fading of a MIMO channel is stronger than a single channel. Multiplexing technologies of MIMO are promising ways to enhance the degree of freedom of main channels. And AN signals can interfere with eavesdroppers while they are sent into the null spaces of desired directions.

We briefly summarize MIMO-based physical layer security techniques of recent research in four categories, as in Fig. 1, including secure unitary beamforming, zero-forcing (ZF) beamforming, convex optimization-based precoding, and AN-based precoding. It is emphasized that a MIMO-based physical layer security scheme of cellular networks can combine multiple techniques.

As in Fig. 1a, unitary beamforming techniques use unitary matrices, semi-unitary matrices, or unitary vectors as transmit precoding, which is similar with traditional beamforming technologies of cellular networks. Unitary beamforming techniques send a message stream via multi-antennas to make it as close to the main channel direction as possible. ZF beamforming is shown in Fig. 1b, which is a secure beamforming technology based on ZF precoding, where the message stream is transmitted to a desired receiver via a shifted beamforming direction, which is orthogonal to the eavesdropper's channel. The illustrative diagram of CVX-based precoding can be seen in Fig. 1c. The problem of optimizing secrecy capacities is usually non-convex in MIMO systems, but Newton methods or Lagrangian dual transformations can be used to trap it in a local maximum problem and address the transmit covariance matrix optimization based on convex optimization tools. AN signals can be transmitted to the null spaces of desired directions, so that they do not affect the desired user, while an eavesdropper's channel is degraded with a high probability, as in Fig. 1d. This AN-based precoding exploits a fraction of the transmit power to send artificially generated noise signals, so the power for transmitting messages will be reduced.

AN-based schemes have been seen as promising methods because this type of techniques has two advantages. Firstly, there are no requirements on the condition of better main channels. Secondly, when the number of transmit antennas is larger the number of the eavesdroppers, the main channel state information (CSI) and precoding matrices of security schemes can be broadcasted to both legitimate receivers and eavesdroppers. We do not need to worry about the



Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks, Fig. 1 Illustrative diagrams of four basic secure multi-antenna technologies. For simplicity, this sets two-dimensional diagrams and a

single message stream. (a) Secure unitary beamforming, (b) Zero-forcing precoding, (c) CVX-based precoding, (d) AN-based precoding (Liu et al. 2017a)

leakage of key precoding messages. More details of the second advantage can be seen in Liu et al. (2017b).

Artificial Noise Schemes

The section will introduce a general AN-based model, which is first provided in Liu et al. (2017b). Assuming t is the number of transmit antennas, messages are encoded in s (which is a variable) strongest eigen-subchannels based on ordered eigenvalues of Wishart matrices ($\mathbf{H}\mathbf{H}^\dagger$ or $\mathbf{H}^\dagger\mathbf{H}$, here, $\mathbf{H}$ is the main channel CSI matrix), while AN signals are generated in remaining $t - s$ eigen-subchannels. This scheme treats the number of eigen-subchannels for message streams, i.e., s , as an optimization objective that can be leveraged to optimize secrecy capacities.

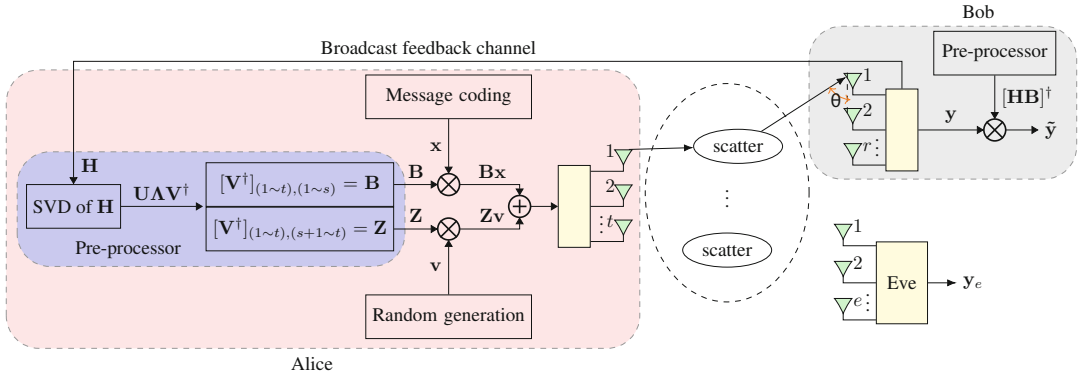
General AN-Based Model

Let us consider a MIMO communication system in the presence of correlated Rayleigh fading at receiver and eavesdropper sides. The system consists of a transmitter (Alice) with t transmit antennas, a legitimate receiver (Bob) with r receive antennas, and an eavesdropper (Eve) with e receive antennas, as shown in Fig. 2. Define $m = \max(t, r)$ and $n = \min(t, r)$. In general, the main channel between Alice and Bob and the wiretap channel between Alice and Eve are defined by receiver-side correlated complex Gaussian matrices $\mathbf{H} \in \mathbb{C}^{r \times t}$ and $\mathbf{H}_e \in \mathbb{C}^{e \times t}$,

whose elements obey distribution $\mathcal{CN}(0, 1)$. $\mathbf{R}_r \in \mathbb{C}^{r \times r}$ and $\mathbf{R}_e \in \mathbb{C}^{e \times e}$ are the receiver-side correlated channel matrices from Bob and Eve, respectively. Also assume that Alice knows full CSI of Bob via a broadcast feedback channel from Bob, including $\mathbf{H}$ and $\mathbf{R}_r$, but only knows the channel distribution information (CDI) of Eve and $\mathbf{R}_e$. Eve knows the CSI of all channels, including $\mathbf{H}$, $\mathbf{H}_e$, $\mathbf{R}_r$, and $\mathbf{R}_e$, where it is assumed that $t > e$.

In this AN-based scheme, there are s ($s \leq t$) message-sending eigen-subchannels, which are selected by Alice based on the CSI feedback from Bob. More specifically, Alice performs SVD of $\mathbf{H}^\dagger\mathbf{H} \in \mathbb{C}^{t \times t}$ in a preprocessor, whose output is a unitary matrix $\mathbf{U} \in \mathbb{C}^{t \times t}$, its Hermitian transpose form $\mathbf{U}^\dagger \in \mathbb{C}^{t \times t}$, and a diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{t \times t}$, which consists of positive and zero eigenvalues of $\mathbf{H}^\dagger\mathbf{H}$. Then, Alice generates a message precoding matrix $\mathbf{B} \in \mathbb{C}^{t \times s}$, whose columns are the eigenvectors corresponding to the first to the s th largest eigenvalues of $\mathbf{H}^\dagger\mathbf{H}$, and an AN precoding matrix $\mathbf{Z} \in \mathbb{C}^{t \times d}$ ($s + d = t$), whose columns are the eigenvectors of the remaining eigenvalues of $\mathbf{H}^\dagger\mathbf{H}$.

Alice transmits $\mathbf{B}\mathbf{w} + \mathbf{Z}\mathbf{v}$ via t transmit antennas. It means that each antenna transmits a combination of message components and AN components, but the AN components can be eliminated by the preprocessor at Bob. In this way, we create a capacity difference between the main channels and wiretap channels. Note that $\mathbf{B}$ and $\mathbf{Z}$ are fixed semi-unitary matrices derived from $\mathbf{H}$. Hence, we have $\mathbf{B}^\dagger\mathbf{B} = \mathbf{I}_s$ and



Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks, Fig. 2 Illustration of an artificial noisy MIMO wiretap channel model with receiver-side correlated Rayleigh fading. Alice has t transmit antennas, Bob has r receive antennas, and Eve has e

receive antennas. Assumed that Alice, Bob, and Eve have uniformly linear array antennas with d_u antenna spacing. θ represents AoA between a scattered path and the antenna array, which can be viewed as a random variable with enough scatterers, and θ follows a Gaussian distribution

$$\text{SVD}(\mathbf{H}^\dagger \mathbf{H}) = \begin{matrix} \boxed{\mathbf{U} : t \text{ rows } t \text{ columns}} \\ \begin{pmatrix} \textcircled{u_{11}} & \textcircled{u_{12}} & \textcircled{u_{13}} & \dots & \textcircled{u_{1t}} \\ \textcircled{u_{21}} & \textcircled{u_{22}} & \textcircled{u_{23}} & \dots & \textcircled{u_{2t}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \textcircled{u_{t1}} & \textcircled{u_{t2}} & \textcircled{u_{t3}} & \dots & \textcircled{u_{tt}} \end{pmatrix} \end{matrix} \times \begin{matrix} \boxed{\mathbf{\Lambda} : t \text{ rows } t \text{ columns}} \\ \begin{pmatrix} \textcircled{\lambda_{11}} & 0 & \dots & 0 \\ 0 & \textcircled{\lambda_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \textcircled{\lambda_{tt}} \end{pmatrix} \end{matrix} \times \mathbf{U}^\dagger$$

for messages
 $\mathbf{B} = \{\mathbf{u}_1, \mathbf{u}_2\}$

for AN
 $\mathbf{Z} = \{\mathbf{u}_3, \dots, \mathbf{u}_t\}$

- 1) The $r \times t$ matrix $\mathbf{H}$ is a channel matrix with t transmit antennas and r receive antennas.
- 2) Two antenna arrays $\{\mathbf{u}_1, \mathbf{u}_2\}$ corresponding to two maximum eigenvalues λ_{11} and λ_{22} are allocated for message transmission.
- 3) Remained antenna array $\mathbf{u}_3, \dots, \mathbf{u}_t$ corresponding to the remained eigenvalues is allocated for AN signal transmission.
- 4) By the precoding $\mathbf{B}$ and $\mathbf{Z}$, the information signal $\mathbf{w}$ and AN signal $\mathbf{s}$ are encoded as $\mathbf{B}\mathbf{w}$ and $\mathbf{Z}\mathbf{s}$.

Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks, Fig. 3 An example of this AN precoding scheme with $s = 2$

$\mathbf{Z}^\dagger \mathbf{Z} = \mathbf{I}_d$. An example of AN precoding scheme with $s = 2$ is illustrated in Fig. 3.

Theorem 1 *The AN signal can be eliminated at Bob if and only if $[\mathbf{H}\mathbf{B}]^\dagger \mathbf{H}\mathbf{Z} = \mathbf{0}$ (Liu et al. 2017b).*

Theorem 2 *The AN signal cannot be eliminated at Eve if and only if $t > e$, $[\mathbf{H}\mathbf{B}]^\dagger \mathbf{H}_e \mathbf{Z} \neq \mathbf{0}$, and $[\mathbf{H}_e \mathbf{B}]^\dagger \mathbf{H}_e \mathbf{Z} \neq \mathbf{0}$ (Liu et al. 2017b).*

According to CSI matrix $\mathbf{H}$ and the preprocessing method, the received signals at Bob and

Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks, Table 1 Secrecy metrics with their conditions

Measures	Symbols	Conditions
Instantaneous secrecy capacity	C_s	$\mathbf{H}_e$ is available and optimal input distribution.
Average secrecy capacity	$\tilde{C}_s$	$\mathbf{H}_e$ is unavailable and optimal input distribution
Instantaneous secrecy rate	R_s	$\mathbf{H}_e$ is available and Gaussian distribution input
Average secrecy rate	$\tilde{R}_s$	$\mathbf{H}_e$ is unavailable and Gaussian distribution input

Eve can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{B}\mathbf{w} + \mathbf{H}\mathbf{Z}\mathbf{v} + \mathbf{n}, \quad (2)$$

$$\mathbf{y}_e = \mathbf{H}_e\mathbf{B}\mathbf{w} + \mathbf{H}_e\mathbf{Z}\mathbf{v} + \mathbf{n}_e, \quad (3)$$

respectively. Here, $\mathbf{w}$ is a transmitted signal of the desired user, and $\mathbf{v}$ is a random AN signal. Both $\mathbf{w}$ and $\mathbf{v}$ are circularly symmetric complex Gaussian vectors with zero means and its covariance matrices $P/t\mathbf{I}_s$ and $P/t\mathbf{I}_d$, respectively, where P is the average transmit power constraint. For simplification, we distribute total power to each antenna evenly. $\mathbf{n}$ and $\mathbf{n}_e$ are additive white Gaussian noise (AWGN) vectors with their covariance matrices $\mathbf{I}_r$ and $\mathbf{I}_e$, respectively.

Bob can eliminate the AN signal $\mathbf{v}$ by preprocessing ($[\mathbf{H}\mathbf{B}]^\dagger \mathbf{H}\mathbf{Z} = \mathbf{0}$) the received signal $\mathbf{y}$ as

$$\tilde{\mathbf{y}} = [\mathbf{H}\mathbf{B}]^\dagger \mathbf{y} = \mathbf{A}_s \mathbf{w} + \tilde{\mathbf{n}}, \quad (4)$$

where $\tilde{\mathbf{n}} = [\mathbf{H}\mathbf{B}]^\dagger \mathbf{n} \in \mathbb{C}^{s \times 1}$ is an AWGN vector with its distribution $\mathcal{CN}(\mathbf{0}, \mathbf{A}_s)$. $\mathbf{A}_s \in \mathbb{R}^{s \times s}$ is a diagonal matrix formed by the first to the s th eigenvalues of $\mathbf{H}^\dagger \mathbf{H}$. In the elimination process, the received signal multiplied by a fixed matrix will not change its capacity. Even if we consider the worst case that Eve has the knowledge of $\mathbf{H}$, $\mathbf{H}_e$, $\mathbf{B}$, and $\mathbf{Z}$, the AN signal still degrades Eve's channel because Eve cannot eliminate the AN signal because $t > e$, $[\mathbf{H}\mathbf{B}]^\dagger \mathbf{H}_e \mathbf{Z} \neq \mathbf{0}$, and $[\mathbf{H}_e \mathbf{B}]^\dagger \mathbf{H}_e \mathbf{Z} \neq \mathbf{0}$, as in **Theorem 1**.

Secrecy Metric

The secrecy capacity and secrecy rate are important measures of AN-based schemes. Here, four expressions are provided to represent the instantaneous secrecy capacity, the average secrecy capacity, the instantaneous secrecy rate, and the average secrecy rate, respectively. These mea-

sures are fit in with different conditions as in Table 1

Instantaneous Secrecy Capacity

In a MIMO wiretap channel model, while Bob has the knowledge of $\mathbf{H}$, $\mathbf{H}_e$, $\mathbf{R}_r$, and $\mathbf{R}_e$, the instantaneous secrecy capacity is

$$C_s = \max_{p(\mathbf{w}), p(\mathbf{v})} \{I(\mathbf{w}; \mathbf{y}) - I(\mathbf{w}; \mathbf{y}_e)\}, \quad (5)$$

where $I(\mathbf{w}; \mathbf{y})$ is the mutual information between information variables $\mathbf{w}$ and received vector $\mathbf{y}$ at Bob and $I(\mathbf{w}; \mathbf{y}_e)$ is the mutual information between information variables $\mathbf{w}$ and received vector $\mathbf{y}_e$ at Eve. And the maximization is taken over all possible input distributions of $p(\mathbf{w})$ and $p(\mathbf{v})$. However, it is hard to begin the optimization process when $\mathbf{H}_e$ is unavailable.

Average Secrecy Capacity

Assuming that the communication lasts longer enough to experience all channel states, to average out the randomness of C_s when only CDI of $\mathbf{H}_e$ is available, we remark that C_s is a function of $\mathbf{H}_e$, and then, the average secrecy capacity is

$$\tilde{C}_s = \max_{p(\mathbf{w}), p(\mathbf{v})} \{I(\mathbf{w}; \mathbf{y}) - I(\mathbf{w}; \mathbf{y}_e | \mathbf{H}_e)\}, \quad (6)$$

where $I(\mathbf{w}; \mathbf{y} | \mathbf{H}_e) = \mathbb{E}_{\mathbf{H}_e}[I(\mathbf{w}; \mathbf{y}_e)]$ is the expected value of the conditional mutual information between $\mathbf{w}$ and $\mathbf{y}_e$ for given $\mathbf{H}_e$. And the maximization is taken over all possible input distributions of $p(\mathbf{w})$ and $p(\mathbf{v})$.

Instantaneous Secrecy Rate

Either in the instantaneous expression or in the average expression, it is hard to find optimal distributions of $p(\mathbf{w})$ and $p(\mathbf{v})$ to maximize the

secrecy capacity. Here follow the convention in Liu et al. (2017b, 2015), and use the secrecy rate instead of the secrecy capacity with Gaussian input alphabets and Gaussian AN, i.e., both $\mathbf{w}$ and $\mathbf{v}$ are circularly symmetric complex Gaussian vectors. In this case, the instantaneous secrecy rate can be expressed as

$$R_s = [C_m - C_w]^+, \quad (7)$$

where $[x]^+ = \max(x, 0)$. C_m and C_w are

$$C_m = \log_2 \det(\mathbf{I}_r + (P/t)\mathbf{H}_1\mathbf{H}_1^\dagger), \quad (8)$$

$$C_w = \log_2 \det\left(\mathbf{I}_e + \frac{(P/t)\mathbf{H}_2\mathbf{H}_2^\dagger}{(P/t)\mathbf{H}_3\mathbf{H}_3^\dagger + \mathbf{I}_e}\right),$$

respectively. Here, $\mathbf{H}_1 = \mathbf{H}\mathbf{B} \in \mathbb{C}^{r \times s}$, $\mathbf{H}_2 = \mathbf{H}_e\mathbf{B} \in \mathbb{C}^{e \times s}$, and $\mathbf{H}_3 = \mathbf{H}_e\mathbf{Z} \in \mathbb{C}^{e \times d}$. However, the instantaneous secrecy capacity is hard to be calculated by Eq. (7) in the absence of Eve's CSI.

Average Secrecy Rate

Assuming the CDI of $\mathbf{H}_e$, the average secrecy rate can be expressed as

$$\begin{aligned} \tilde{R}_s &= \mathbb{E}_{\mathbf{H}_e} [C_m - C_w]^+ \\ &\geq C_m - \mathbb{E}_{\mathbf{H}_e} [C_w]. \end{aligned} \quad (9)$$

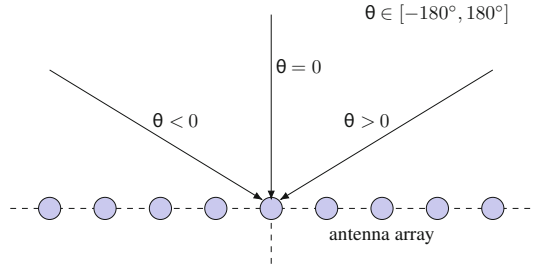
Equation (9) gets an equality if and only if the secrecy rates are always positive over all channel states. Since AN signals are used to jam the wiretap channels, the average secrecy rate is always nonnegative. From **Lemma 1** in the Appendix, $\mathbf{H}_2$, $\mathbf{H}_3$, and $\mathbf{H}_4$ are complex Gaussian matrices with their distributions

$$\mathbf{H}_2 \sim \mathcal{CN}_{e,s}(\mathbf{0}, \mathbf{R}_e \otimes \mathbf{I}_s), \quad (10)$$

and

$$\mathbf{H}_3 \sim \mathcal{CN}_{e,d}(\mathbf{0}, \mathbf{R}_e \otimes \mathbf{I}_d), \quad (11)$$

respectively. Hence, the expected value of average secrecy rates can be calculated by Monte Carlo simulations with Eve's CDI.



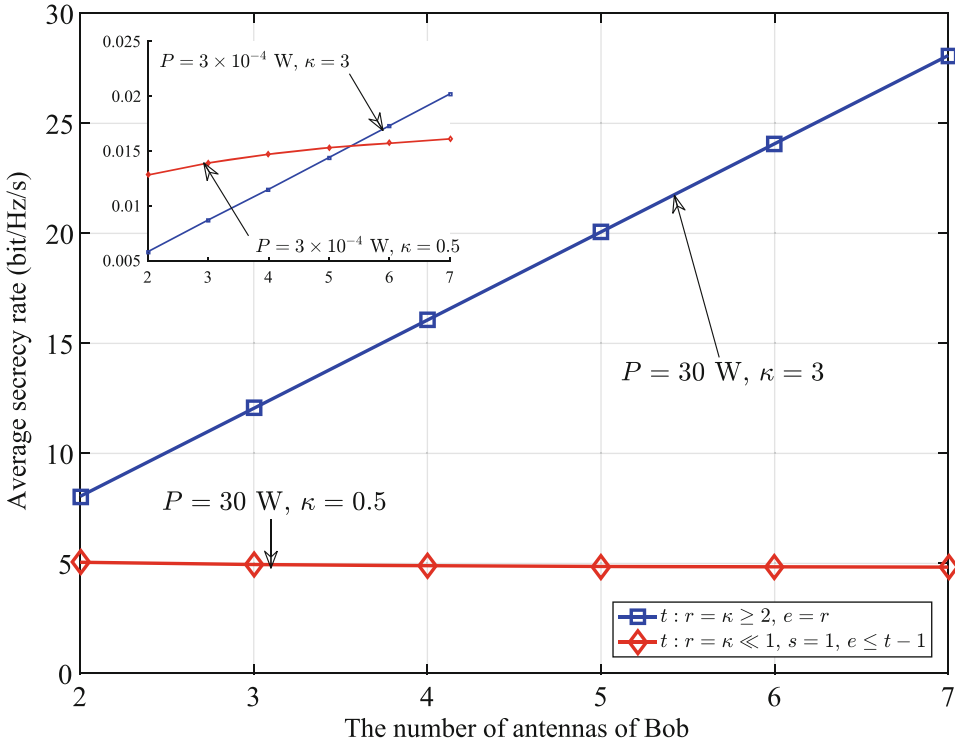
Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks, Fig. 4 AoA model with respect to a uniform antenna array

Correlated Matrices

The correlated matrices $\mathbf{R}_r$ and $\mathbf{R}_e$ can affect all the secrecy metrics of an AN-based system. As in Bolcskei et al. (2003), a correlated matrix $\mathbf{R}$ (generalized for $\mathbf{R}_r$ and $\mathbf{R}_e$) is a function of AoA (defined by θ) distribution. Here consider an AoA model with respect to a uniform antenna array based on 3GPP 3GP (2003), as shown in Fig. 4, where the AoA model belongs to a Gaussian distribution, the mean AoA of θ is $\bar{\theta}$, and the RAS (variance) of θ is δ . Then, we get the following simple model for the correlated coefficient $[\mathbf{R}]_{u,v}$, which is

$$\begin{aligned} [\mathbf{R}]_{u,v} &= \exp \left\{ -j2\pi(u-v)d_u \cos \bar{\theta} \right\} \\ &\times \exp \left\{ -\frac{1}{2}(2\pi\delta(u-v)d_u \sin \bar{\theta})^2 \right\}, \end{aligned} \quad (12)$$

where $u \in \{1, \dots, r\}$ and $v \in \{1, \dots, r\}$ are the receive antenna index numbers. According to the parameters from 3GPP standard 3GP (2003), $\bar{\theta}$ is set in the range $[0^\circ, 100^\circ]$ and δ in the range $[5^\circ, 35^\circ]$, respectively. Assume that all antennas form a uniformly linear antenna array with $d_u = d_{\min}/\omega$, where $d_{\min}$ is the normalized minimum distance and ω is the wavelength. In the wiretap channel model, $\mathbf{R}_r$ and $\mathbf{R}_e$ have the same structure as $\mathbf{R}$. Here define the mean AoA at Bob and Eve as $\bar{\theta}_r$ and $\bar{\theta}_e$, respectively, and define the RAS at Bob and Eve as δ_r and δ_e , respectively.



Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks, Fig. 5 Average secrecy rates in terms of the number of Bob's antennas

Simulations

Simulation results are provided to investigate joint impacts of the number of antennas and the number of selected message-sending eigen-subchannels on the average secrecy rates.

Figure 5 shows the relationship between the average secrecy rates and the number of Bob's antennas. In the case of adequate transmit antennas, i.e., $t:r = \kappa \geq 2$ and $e = r$, the average secrecy rates increase with an increasing number of Bob's antennas in both high ($P = 30$ W) and low SNR ($P = 3 \times 10^{-4}$ W) regions. In the case with a small number of transmit antennas and adequate receive antennas (here we set $t:r = \leq 1, s = 1$, and $e = t - 1$), we find that the average secrecy rates converge to a deterministic constant when e becomes large. It means that an increasing number of Bob's antennas bring in no benefit when Alice uses less antennas than Bob

and selects $s = 1$, i.e., uses a transmit diversity scheme.

Conclusion and Further Work

Physical layer security will play a critical role in the future security architecture of cellular communications. This research should keep competing with cryptography and quantum communications, which are believed to be the three major security architectures for cellular communication systems. AN-based schemes are seen as promising methods and now have produced a raft of fascinating and important results. However, these AN-based schemes are optimal only under the condition of that the number of transmit antennas is larger than eavesdropper antennas. When the number of transmitted antennas is constrained or even smaller than that of eavesdropper

antennas, AN-based schemes cannot get positive secrecy capacities or rates. This motivates us to design a better AN-based scheme. In addition, the secrecy capacity (rate) quantization is a great challenge in AN-based schemes. Without perfect CSI of eavesdroppers, it is hard for encoders to calculate instantaneous secrecy capacities or rates. And the influence of fading does not allow us to use a simple AWGN or Rayleigh channel model to calculate a secrecy rate. It seems that we need to investigate main and wiretap fading channels that belong to other models of cellular communications. At last, power allocation in cellular networks should be investigated further, because power allocation problems are usually non-convex ones with a lot of cellular users.

Appendix

Lemma 1 (Proved in Gupta and Nagar 1999) Define an $r \times t$ matrix $\mathbf{A} \sim \mathcal{CN}_{r,t}(\mathbf{0}, \mathbf{R} \otimes \mathbf{I}_t)$ as a receiver-side correlated central complex Gaussian matrix, and establish an independent $(t \times s)$ unitary matrix $\mathbf{B}$. We have

$$\mathbf{AB} \sim \mathcal{CN}_{r,s}(\mathbf{0}, \mathbf{R} \otimes \mathbf{I}_s), \quad (13)$$

where $s \in \mathbb{R}$ and $t \geq s$.

Key Applications

Artificial noise (AN) schemes based on MIMO technologies can be applied to various cellular scenarios. For example, Massive MIMO systems have an enormous number of antennas, which offer more degrees of freedom for wireless channels, and a more secure performance in terms of secrecy capacities and the number of AN beam-forming. The small cell base stations deployed as cooperative jammers in cellular networks can be used to provide well-designed AN signals. In addition, the long-term evolution advanced (LTE-A) system supports device to device (D2D) communications, which is defined as the direct communications between two mobile users via

shared radio resources with cellular users. D2D interference caused by the shared radio resources can be seen as AN signals to interfere with the illegitimate eavesdropper.

Cross-References

- IoT Security
- Security and Privacy in 4G/LTE Network
- Vehicle Ad-Hoc Networks: Services, Security, and Privacy
- Wireless Data Security
- Wireless Jamming Attack
- Wireless Key Establishment
- Wireless Sensor Network Security

References

- 3GP (2003) Spatial channel model for multiple input multiple output (MIMO) simulations. Rev. 6
- Bolskei H, Borgmann M, Paulraj AJ (2003) Impact of the propagation environment on the performance of space-frequency coded MIMO-OFDM. *IEEE J Sel Areas Commun* 21(3):427–439. <https://doi.org/10.1109/JSAC.2003.809723>
- Gupta AK, Nagar DK (1999) Matrix variate distributions, vol 104. CRC Press, London
- Liu S, Hong Y, Viterbo E (2015) Artificial noise revisited. *IEEE Trans Inf Theory* 61(7):3901–3911. <https://doi.org/10.1109/TIT.2015.2437882>
- Liu Y, Chen HH, Wang L (2017a) Physical layer security for next generation wireless networks: theories, technologies, and challenges. *IEEE Commun Surv Tutorials* 19(1):347–376. <https://doi.org/10.1109/COMST.2016.2598968>
- Liu Y, Chen HH, Wang L (2017b) Secrecy capacity analysis of artificial noisy MIMO channels—an approach based on ordered eigenvalues of Wishart matrices. *IEEE Trans Inf Forensics Secur* 12(3):617–630. <https://doi.org/10.1109/TIFS.2016.2627219>
- Wyner AD (1975) The wire-tap channel. *Bell System Tech J* 54(8):1355–1367

Asynchronous Coordination Scheme/Protocol

- Asynchronous Coordination Techniques

Asynchronous Coordination Techniques

Ruoyu Su¹, Dengyin Zhang¹, Ramachandran Venkatesan², Cheng Li², Zijun Gong³, and Fan Jiang³

¹School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China

²Faculty of Engineering and Applied Science, Memorial University, St. John's, NL, Canada

³Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

Synonyms

Asynchronous coordination scheme/protocol; Asynchronous wake-up coordination scheme/protocol; Asynchronous wake-up scheme/protocol

Definition

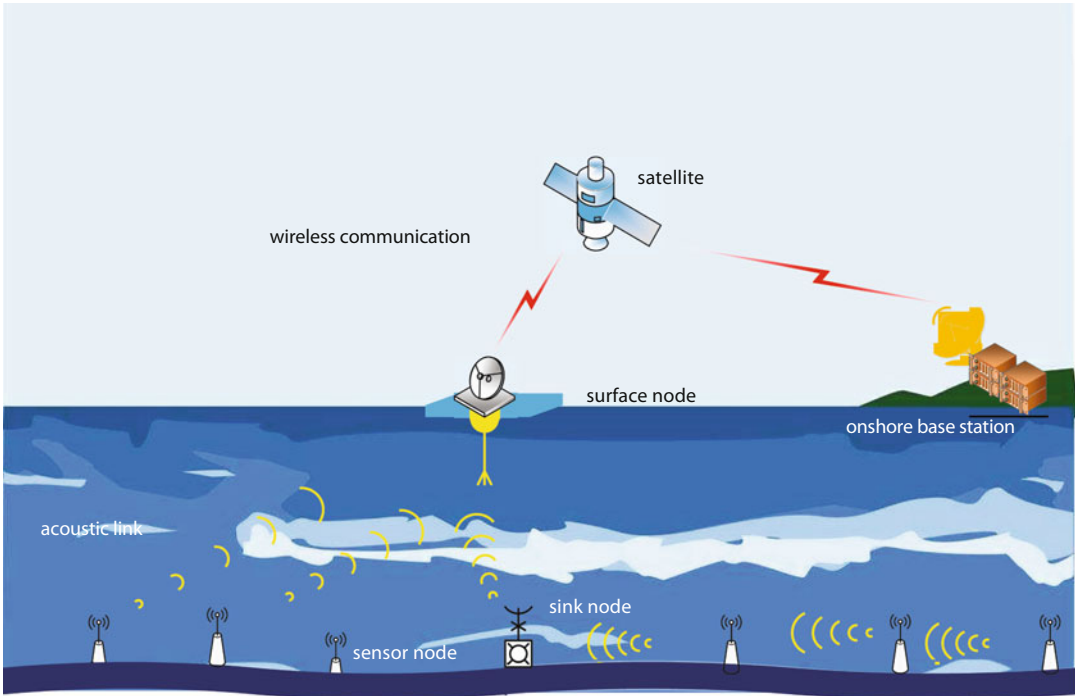
Asynchronous coordination technique is a communication approach that sensor nodes use different predetermined wake-up and sleep patterns for each cycle of network operations in order to achieve high energy efficiency and to guarantee network connectivity based on certain properties in linear algebra and optimization. No time synchronization is required in the network operations.

Historical Background

Underwater acoustic sensor networks (UWSNs) have attracted much research interest in recent years due to their wide range of potential applications such as marine environmental monitoring, undersea resources exploration, disaster monitoring and prevention, assisted location and navigation, and security monitoring

(Akyildiz et al. 2015). Early generation of UWSNs include the Acoustic Local Area Network (ALAN), which was deployed by the Woods Hole Oceanographic Institution (WHOI) in 1994 off the coast of Monterey, California (Catipovic et al. 1993). ALAN, aiming at long-term data acquisition and ocean monitoring, consists of a number of sensor nodes placed on the seafloor and one surface node attached to a buoy. The underwater sensor nodes located on the seafloor transmit data to the surface node via acoustic links. Data was transmitted to the onshore base station via a radio frequency link. Another example of UWSNs is Seaweb 2000 (Rice et al. 2000), which was one of the first multi-hop underwater acoustic networks. This project advanced the underwater communication technology to achieve near-realtime, synoptic observation of the underwater environment. In Seaweb 2000, 14 sensor nodes (to sense, collect, and relay data) were placed on the seafloor and three gateway nodes with buoys (to collect data from sensor nodes and transmit data to the onshore station) were deployed on the sea surface (Rice et al. 2002). A fundamental architecture of UWSNs is presented in Fig. 1. As shown in Fig. 1, underwater acoustic sensor nodes either directly report data to the surface node or transmit data to the sink node via multi-hop. The sink node aggregates data from different sensor nodes and then transmits to the surface node. The onshore base station will receive data from the satellite after the surface node reports collected data to the satellite.

Underwater acoustic communications consume more energy especially for long-range transmission (Zhang et al. 2015) compared with terrestrial communications. Furthermore, energy availability is limited in underwater acoustic communications because of the unchangeable and nonchargeable batteries in the sensor nodes. Frequently replacing sensor nodes generates high deployment cost, which is not realistic in long-term marine monitoring applications (Heidemann et al. 2012; Zhang et al. 2015). Therefore, unlike the high requirements of network throughput and packet delay, energy efficiency is a fundamental



Asynchronous Coordination Techniques, Fig. 1 The fundamental architecture of UWSNs

and significant issue when designing underwater network protocols.

An effective approach to energy saving is the wake-up coordination scheme. In the wake-up coordination scheme, each sensor node works periodically and has two modes of operations in each period: wake-up mode and sleep mode. In the wake-up mode, the transceiver is turned on for neighboring node discovery and data exchange (once the communication between a source node and destination node is established), whereas it remains off during the sleep mode so that energy consumption can be reduced. In general, the wake-up coordination scheme can be categorized into on-demand, synchronous, and asynchronous wake-up coordination schemes.

On-demand Wake-up Coordination Schemes

On-demand wake-up coordination scheme is a communication approach that each sensor node

is equipped with a low power radio module to wake up the node for communication. The wake-up time can be predetermined according to application requirements, network data traffic, and so on. For example, cooperative underwater multichannel media access control (MAC) protocol (CUMAC) is a typical protocol based on the on-demand wake-up coordination mechanism, which adopts two acoustic transceivers to be equipped for one sensor node (Zhou et al. 2012). The low-priced acoustic transceiver is used to detect the channel by sending and receiving tone signal. The other acoustic transceiver is for data transmission. The tone pulse sent by the low-price acoustic transceiver at a predetermined time arrives at different sensor nodes at different times. Sensor nodes can catch the tone pulse by the low-price acoustic transceiver at the right time and then wake up the acoustic transceiver to exchange data packets at proper time. Similar mechanisms can be referred in Syed et al. (2008) and Jurdak et al. (2010).

Synchronous Wake-up Coordination Schemes

Synchronous wake-up coordination scheme is a communication approach that all sensor nodes are time synchronized before data transmission so that each sensor node can wake up at the proper time to send or receive data. Also, sensor nodes can turn off their transceivers to conserve energy when there is no data to transmit or receive. For example, underwater wireless acoustic networks MAC protocol (UWAN-MAC) is a distributed scalable energy-efficient scheme based on synchronous wake-up coordination mechanism (Min and Rodoplu 2008). In the UWAN-MAC protocol, before data transmission, each sensor node broadcasts signaling packets (i.e., time stamp) which includes information about the amount of time in sleep mode during one cycle and records information transmitted by its neighboring nodes to achieve time synchronization among all sensor nodes. The time stamps guarantee that neighboring nodes would wake up at the correct time in order to receive data packets without the knowledge of the propagation delay. Similar protocols can be found in van Dam and Langendoen (2003) and Ye et al. (2004).

Asynchronous Wake-up Coordination Schemes

Unlike the on-demand and synchronous wake-up coordination schemes, asynchronous wake-up coordination schemes utilize different predetermined wake-up and sleep patterns for each cycle of network operations to achieve high energy efficiency and to guarantee network connectivity without time synchronization during the network operations. By using asynchronous wake-up coordination schemes, two sensor nodes are able to communicate with each other at least once in one cycle if they are in the communication range. Meanwhile, sensor nodes can sleep as long as possible to reduce energy consumption when they do not have data to transmit and to forward. In general, each sensor node works periodically. In one period, the operations of

each sensor node alternate between the wake-up mode and sleep mode. In the wake-up mode, the transceiver of the sensor node is turned on for neighboring node discovery and data exchange (once the communication between a source node and destination node is established). In the sleep mode, the transceiver of the sensor node remains off so that energy consumption can be reduced. Both the wake-up mode and the sleep mode may contain different numbers of time slots. A time slot is called an active slot when the sensor node operates in the wake-up mode and is called an inactive slot when the sensor node stays in the sleep mode. In this case, the operation status of a sensor node alternates between a predetermined number of active slots and inactive slots. To understand the diverse asynchronous wake-up coordination schemes, the following definitions and theorems are presented.

Preliminaries

Definition 1 (Quorum system) (Jiang et al. 2005) Let U be a universal set and $U = \{0, 1, 2, \dots, n - 1\}$. A quorum system Q is defined as a collection of non-empty subsets of U (e.g., G and H), which satisfies

$$\forall G, H \in Q, G \cap H \neq \emptyset. \quad (1)$$

For example, $U = \{0, 1, 2\}$, $Q = \{\{0, 1\}, \{0, 2\}, \{1, 2\}\}$ is a quorum system under U .

Definition 2 (Rotational closure property) (Jiang et al. 2005) A quorum system has the rotation closure property when

$$\begin{aligned} \forall G, H \in Q, i \in \{0, 1, 2, \dots, n - 1\}, \\ G \cap \text{rotate}(H, i) = \emptyset, \end{aligned} \quad (2)$$

where i is a nonnegative integer and $\text{rotate}(H, i) = \{(j + i) \bmod n | j \in H\}$.

For example, the quorum system $Q = \{\{0, 1\}, \{0, 2\}, \{1, 2\}\}$ under $\{0, 1, 2\}$ has the rotation closure property. However, the quorum system $Q' = \{\{0, 1\}, \{0, 2\}, \{0, 3\}, \{1, 2, 3\}\}$ under $U = \{0, 1, 2, 3\}$ has no rotation closure property

because $\{0,1\} \cap \text{rotate}(\{0,3\}, 3) = \emptyset$ (Jiang et al. 2005).

Definition 3 ($(Z_v, +)$ (Stinson 2004) $(Z_v, +)$ is a finite group of order v , where v is a positive integer.

Definition 4 (Cyclic difference set) (Stinson 2004; Choi and Shen 2011) Given a set D with v elements and k subset elements, $D : a_1, a_2, \dots, a_k \pmod{v}$ is called a (v, k, λ) –cyclic difference set in $(Z_v, +)$ if for every $d \neq 0 \pmod{v}$ there are exactly λ ordered pairs (a_i, a_j) , such that $a_i - a_j = d \pmod{v}$.

For example, $D = \{1, 2, 4\}$ is called a $(7, 3, 1)$ –cyclic difference set in $(Z_7, +)$.

Theorem 1

(Singer difference set) (Stinson 2004 ; Choi and Shen 2011) Let q be a prime power. Then there exists a $(q^2 + q + 1, q + 1, 1)$ –cyclic difference set in $(Z_{q^2+q+1}, +)$.

The cyclic difference sets can be constructed according to Theorem 1. For instance, $(7, 3, 1)$ –cyclic difference set when $q = 2$, $(13, 4, 1)$ –cyclic difference set when $q = 3$, $(21, 5, 1)$ –cyclic difference set when $q = 4$. More cyclic difference sets can be found in Appendix A of Stinson (2004).

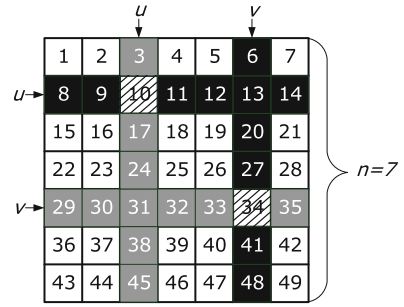
Theorem 2

(Multiplier theorem) (Stinson 2004) Suppose there exists a (v, k, λ) –difference set D in an Abelian group $(Z_v, +)$ of order v . p is a multiplier of D if the following four conditions are satisfied: (1) p is prime; (2) $\gcd(p, v) = 1$, $\gcd$ is greatest common divisor; (3) $k - \lambda = 0 \pmod{p}$; and (4) $p > \lambda$.

According to Theorems 1 and 2, $p = 2$ is a multiplier of $(21, 5, 1)$ –cyclic difference set because 2 satisfies the four conditions in Theorem 2. Based on the multiplier, it is easy to construct two $(7, 3, 1)$ –cyclic difference sets: $\{1, 2, 4\}$ and $\{3, 5, 6\}$.

Quorum-Based Energy Conserving (QEC) Protocol

Figure 2 presents the basic idea of network operations in QEC. The white cube represents an inactive

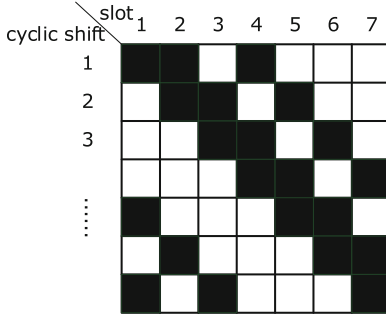


Asynchronous Coordination Techniques, Fig. 2 The fundamental idea of QEC

tive slot and the black and gray cube are an active slot. When node u and node v each arbitrarily choose a row and a column from a square of size n , respectively, their active intervals would overlap for at least two active slots, e.g., the 10-th and the 34-th slots shown in this figure ($n = 7$). Furthermore, there exists at least one overlapped active slot if two sensor nodes adopt two different grid sizes (Chao et al. 2006). It is obvious that a large value of n can lead to better energy conservation (larger number of time slots in one period associated with larger number of inactive slots) but it would also increase the latency. The QEC (Chao et al. 2006) accommodates different traffic loads by varying the value of n and thus achieve a balance between energy efficiency and packet delay. The simulation results show that QEC can outperform IEEE 802.11 in terms of energy consumption.

Cyclic Difference Set (CDS)-Based Protocol

The CDS-based protocol determines the number and positions of active and inactive slots in one cycle using as Singer difference set of a (v, k, λ) –cyclic difference set, where v , k , and λ represent the total number of slots, the number of active slots, and the minimum number of overlapping active slots in one cycle, respectively (Zheng et al. 2003, 2006). The CDS-based protocol guarantees at least one active overlapping active slot between any two sensor nodes for any cyclic shift.



Asynchronous Coordination Techniques, Fig. 3 An example of the network operations in the CDS-based protocol followed (7,3,1)–cyclic different set

Figure 3 presents an asynchronous wake-up coordination schemes followed (7,3,1)–cyclic different set. In one period, a sensor node has seven time slots including three active slots and four inactive slots. The positions of active slots (i.e., black cubes) and inactive slots (i.e., white cubes) are determined by the (7,3,1)–cyclic different set. There exists at least one active overlapping slot between any two cycle shift for the cases of perfect and no alignment of time slot boundaries Zheng et al. (2006). That is, the network connectivity can be guaranteed and the energy consumption on idle listening is reduced as well. Furthermore, the CDS-based protocol conserves more energy on idle listening than that of the QEC protocol due to the minimum number of overlapping active slots is one between any cycle shift.

Extension of CDS-Based Protocols

The extension of CDS-based protocols aims to reduce more energy consumption by leveraging a longer cycle period with less active slots without sacrificing network connectivity.

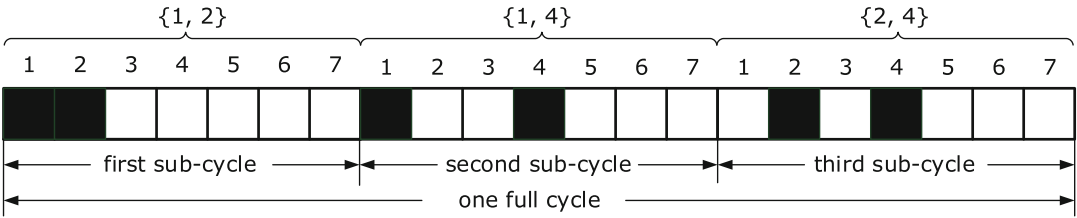
Exponential adaptive CDS-based protocol (EACDS) and multiplicative adaptive CDS-based protocol (MACDS) are the two extensions of the CDS-based protocols (Choi and Shen 2011). The key idea of these two schemes is the use of hierarchical arrangements of sets with the Kronecker product (Anderson 1998; Bernstein 2008). The EACDS-based scheme utilized a difference set scaled by another difference

set, which is called the exponential set. The scaling was done by the Kronecker product and guaranteed that there was at least one overlapping slot between any two EACDSs with different duty cycles. For example, a difference set (v_I, k_I, λ_I) , called an initial set, is scaled by another difference set (v_E, k_E, λ_E) , called an exponential set. Then, the higher level hierarchical set can be obtained by the Kronecker product, i.e., $(v_I, k_I, \lambda_I) \otimes (v_E, k_E, \lambda_E)$, where operator $\otimes$ denotes the Kronecker product. The MACDS-based scheme used a multiplier set instead of the exponential set. All multiplier sets were selected from the relaxed difference sets (Luk and Wong 1997). The performance evaluation revealed that the MACDS-based scheme conserves more energy, whereas the EACDS is more suitable for scenarios where many different duty cycles are required.

Another extension of CDS-based protocol is to utilize permutation and combination to generate a longer cycle period for each sensor node (Su et al. 2016). Based on the Singer difference set, there exist $q + 1$ $\binom{q}{q+1}$ possible position combinations for q active slots in one subcycle. For example, when $q = 2$, we get a (7,3,1)–cyclic different set: $D = \{1,2,4\}$. There are three possible position combinations which has two active slots in one subcycle, and they are $\{1,2\}$, $\{1,4\}$, and $\{2,4\}$, as shown in Fig. 4.

One full operating cycle of a sensor node contains concatenation of all possible subcycles, each corresponding to a distinct position combination of any q active slots. Therefore, an operating cycle is comprised of $(q + 1)$ subcycles, which corresponds to $(q + 1) \cdot (q^2 + q + 1)$ slots, $(q + 1) \cdot q$ of which are active slots. Any ordering of subcycles can be used for concatenation. One working cycle period of a sensor node contains $(q + 1) \cdot (q^2 + q + 1)$ active slots. Thus, the duty cycle is $q/(q^2 + q + 1)$. Figure 4 shows an example when $q = 2$.

Table 1 summarizes duty cycles used in Su et al. (2016) and the CDS-based protocol (Zheng et al. 2003). For a given value of q , the proposed scheme in Su et al. (2016) can generate a smaller duty cycle than the CDS-based protocol. It is obvious that a larger value of q can achieve a



Asynchronous Coordination Techniques, Fig. 4 Extension of the CDS-based protocol by permutation and combination

Asynchronous Coordination Techniques, Table 1 Duty cycles used in Su et al. (2016) and the CDS-based protocol (Zheng et al. 2003)

q	Duty cycles (%) (Su et al. 2016)	Duty cycles (%) (Zheng et al. 2003)
2	28.57	42.85
3	23.08	30.77
4	19.05	23.81
5	16.13	19.35
7	12.28	14.04
8	10.96	12.33
9	9.89	10.99
11	8.27	9.02

smaller duty cycle and thus reduce more energy consumption for idle listening. However, normally, for a given smaller duty cycle, a sensor node requires more active slots to connect to a next-hop sensor node.

Tables 2 and 3 compare the proposed scheme in Su et al. (2016) that is extended using Kronecker product to the EACDS and MACDS proposed in Choi and Shen (2011), respectively. For EACDS, $(v_E, k_E, \lambda_E) = (3, 2, 1)$. For MACDS, (v_M, k_M, λ_M) is selected from the rotational set group as used in Luk and Wong (1997). For the scheme proposed in Su et al. (2016) and the EACDS and MACDS, $q = 2$ is chosen as an example. From Tables 2 and 3, for given values of q, n , and (v_M, k_M, λ_M) , the smaller duty cycles can be generated by the scheme proposed in Su et al. (2016) using the Kronecker product rather than EACDS and MACDS.

Heterogenous CDS-Based Schemes

The length of a cycle period and the corresponding number of active slots determines the energy

Asynchronous Coordination Techniques, Table 2 Duty cycles used in Su et al. (2016) with Kronecker product and EACDS

n	Duty cycles (%) (Su et al. 2016)	Duty cycles in EACDS (%) (Choi and Shen 2011)
1	19.04	28.57
2	12.70	19.04
3	8.47	12.70
4	5.64	8.47
5	3.76	5.64
6	2.51	3.76

Asynchronous Coordination Techniques, Table 3 Duty cycles used in Su et al. (2016) with Kronecker product and MACDS

(V_M, k_M, λ_M)	Duty cycles (%) (Su et al. 2016)	Duty cycles in MACDS (%) (Choi and Shen 2011)
(4,3,1)	21.43	32.14
(5,3,1)	17.14	25.71
(6,3,1)	14.29	21.42
(12,4,1)	9.52	14.29
(24,6,1)	7.14	10.71
(48,8,1)	4.76	7.14

consumption of underwater sensor nodes when data exchanges do not frequently occur in long-term monitoring applications. When some nodes need to change the cycle length dynamically to satisfy the requirements of energy and packet delay and the other sensor nodes keep unchanged, the network connectivity will be lost when the wake-up coordination scheme do not guarantee that any two sensor nodes can communicate at least once in one cycle period. The heterogeneous CDS-based schemes can keep the network connectivity when different CDS-based wake-up

coordination schemes are adopted by different sensor nodes.

Cyclic quorum system pair (CQS-pair) is a heterogenous asynchronous wake-up coordination scheme based on the CDS-based protocol, which is derived by using the Singer cyclic difference set (Theorem 1) and the multiplier theorem (Theorem 2) (Lai et al. 2010). Moreover, a verification matrix is used to verify that any two heterogenous difference sets have the rotation closure property. To be more specific, sensor node a adopts a difference set $\{a_1, a_2, a_3, \dots, a_k\}$ in $(Z_N, +)$ and sensor node b adopts another difference set $\{b_1, b_2, b_3, \dots, b_l\}$ in $(Z_M, +)$, where $N \leq M$ and $p = \lceil M/N \rceil$. The proposed verification matrix (Lai et al. 2010) is defined as a $pk \times l$ matrix $M_{l \times pk}$ whose element $m_{i,j}$ in the matrix is calculated by $(b_i - a_j^p) \pmod{M}$ and $a_j^p \in A^p$, as presented as follows.

$$M_{l \times pk} = \begin{bmatrix} b_1 - a_1^p & \cdots & b_1 - a_{pk}^p \\ \vdots & b_i - a_j^p & \vdots \\ b_l - a_1^p & \cdots & b_l - a_{pk}^p \end{bmatrix} \quad (3)$$

Sensor node a and b have the rotation closure property if $M_{l \times pk}$ contains all elements from 0 to $M - 1$, which means these two heterogeneous difference sets can keep network connectivity when sensor nodes utilize them respectively.

For example, according to Theorems 1 and 2, there are two (7,3,1)-cyclic difference sets: $\{1,2,4\}$ and $\{3,5,6\}$ in $(Z_7, +)$ when a cycle period contains 7 slots. Similarly, there are two (21,5,1)-cyclic difference sets: $\{3,6,7,12,14\}$ and $\{7,9,14,15,18\}$ in $(Z_{21}, +)$ when a cycle period contains 21 slots. By using Eq. (3), sensor nodes can switch between a (7,3,1)-cyclic difference set with $\{1,2,4\}$ and a (21,5,1)-cyclic difference set with $\{7,9,14,15,18\}$ without sacrificing the network connectivity. Another pair of wake-up coordination schemes is a (7,3,1)-cyclic difference set with $\{3,5,6\}$ and a (21,5,1)-cyclic difference set with $\{3,6,7,12,14\}$. Unlike the proposed schemes in Chao et al. (2006), Zheng et al. (2003), Choi and Shen (2011), and Su et al. (2016), the CQS-pair (Lai et al. 2010) provides more choices

for sensor nodes to adapt dynamic changes of network requirements without further improving the energy consumption for long-term monitoring applications.

Key Applications

Asynchronous wake-up coordination scheme has been involved in different scenarios, such as long-term marine environmental monitoring, undersea resource explorations, assisted location and navigation, delay-tolerant data transmission, and some energy-limited applications. With the development of internet of underwater things (IoUT), the communications among multiple underwater autonomous vehicles (AUV) and the cooperation between AUV and glider can utilize asynchronous wake-up coordination scheme to prolong the network lifetime and extend the monitoring area.

Future Directions

Asynchronous wake-up coordination scheme brings energy efficiency for UWSNs. However, it also prolongs the data packet delay compared with on-demand and synchronous wake-up coordination schemes. Some challenges are presented below.

- Balancing between energy consumption and packet delay

It is obvious that the longer cycle period leads to low duty cycles for sensor nodes so as to significantly improve the energy efficiency. However, the packet delay is prolonged because not all sensor nodes are always active and the neighboring nodes discovery requires more time (or time slots) with a low duty cycle. Consequently, to achieve a balance between packet delay and energy consumption is a challenge for UWSNs when considering the nontrivial propagation delay by adopting asynchronous wake-up coordination scheme. On the other aspect, according to the data traffic or historic connection information (e.g., which active slots always enable the

communication connection), how to select a proper duty cycle (i.e., network operation pattern such as (7,3,1)–cyclic difference set-sets: {1,2,4}) achieves a balance between energy consumption and packet delay, which can be further investigated.

- Mobility of sensor nodes

In traditional UWSNs, all sensor nodes are supposed to be fixed on the seafloor in UWSNs. In the internet of underwater things (IoUT), AUVs serve as mobile sensor nodes, which can gather data in a large area for long-term monitoring applications. Directly utilizing aforementioned asynchronous wake-up coordination schemes may lead to non-trivial energy consumption and packet delay caused by the neighboring node discovery when a mobile sensor node is out of communication range of its intended nodes. How to adjust the asynchronous wake-up coordination scheme to accommodate the varying network topology in an energy-efficient manner is still an open issue in this area.

Cross-References

- [Asynchronous Coordination Scheme/Protocol](#)
- [Asynchronous Wake-Up Coordination Scheme/Protocol](#)
- [Asynchronous Wake-Up Scheme/Protocol](#)

Acknowledgment This work was supported in part by the National Natural Science Foundation of China under Grant 61571241.

References

- Akyildiz F, Wang P, Sun Z (2015) Realizing underwater communication through magnetic induction. *IEEE Commun Mag* 53(11):42–48
- Anderson I (1998) Combinatorial designs and tournaments. Oxford University Press, Oxford
- Bernstein D (2008) Matrix mathematics: theory, facts, and formulas. Princeton University Press, Princeton
- Catipovic J, Brady D, Etchemendy S (1993) Development of underwater acoustic modems and networks. *Oceanography* 6(3):112–119
- Chao C, Sheu J, Chou I (2006) An adaptive quorum-based energy conserving protocol for IEEE 802.11 ad hoc networks. *IEEE Trans Mob Comput* 5(5):560–570
- Choi B, Shen X (2011) Adaptive asynchronous sleep scheduling protocols for delay tolerant networks. *IEEE Trans Mob Comput* 10(9):1283–1296
- Heidemann J, Stojanovic M, Zorzi M (2012) Underwater sensor networks: applications, advances and challenges. *Philos Trans* 370(1958):158
- Jiang J, Tseng Y, Hsu C, Lai T (2005) Quorum-based asynchronous power-saving protocols for IEEE 802.11 ad hoc networks. *Mob Netw Appl* 10(1–2):169–181
- Jurdak R, Ruzzelli A, O'Hare G (2010) Radio sleep mode optimization in wireless sensor networks. *IEEE Trans Mob Comput* 9(7):955–968
- Lai S, Ravindran B, Cho H (2010) Heterogenous quorum-based wake-up scheduling in wireless sensor networks. *IEEE Trans Comput* 59(11):1562–1575
- Luk W, Wong T (1997) Two new quorum based algorithms for distributed mutual exclusion. In: International conference on distributed computing systems, Baltimore, pp 100–106
- Min KP, Rodoplu V (2008) UWAN-MAC: an energy-efficient MAC protocol for underwater acoustic wireless sensor networks. *IEEE J Ocean Eng* 32(3):710–720
- Rice J, Creber B, Fletcher C, Baxley P, Rogers K, McDonald K, Rees D, Wolf M, Merriam S, Mehio R, Proakis J, Scussell K, Porta D, Baker J, Hardiman J, Green D (2000) Evolution of seaweb underwater acoustic networking. In: Proceedings of the MTS/IEEE conference and exhibition for ocean engineering, science and technology (OCEANS), Providence, vol 3, pp 2007–2017
- Rice J, Amundsen K, Scussell K (2002) Seaweb 2002 experiment quick-look report. SPAWAR Systems Center, San Diego
- Stinson D (2004) Combinatorial designs: constructions and analysis. Springer, New York
- Su R, Venkatesan R, Li C (2016) An energy-efficient asynchronous wake-up scheme for underwater acoustic sensor networks. *Wirel Commun Mob Comput* 16(9):1158–1172
- Syed A, Ye W, Heidemann J (2008) Comparison and evaluation of the T-Lohi MAC for underwater acoustic sensor networks. *IEEE J Sel Areas Commun* 26(9):1731–1743
- van Dam T, Langendoen K (2003) An adaptive energy-efficient MAC protocol for wireless sensor networks. In: Proceedings of the ACM conference on embedded networked sensor systems (SenSys), Los Angeles, pp 171–180
- Ye W, Heidemann J, Estrin D (2004) Medium access control with coordinated adaptive sleeping for wireless sensor networks. *IEEE/ACM Trans Networking* 12(3):493–506
- Zhang X, Cui J, Das S, Gerla M (2015) Underwater wireless communications and networks: theory and application: part 1 [guest editorial]. *IEEE Commun Mag* 53(11):40–41
- Zheng R, Hou C, Sha L (2003) Asynchronous wakeup for ad hoc networks. In: Proceedings of the ACM

international symposium on mobile ad hoc networking & computing (MobiHoc), Annapolis, pp 35–45

Zheng R, Hou C, Sha L (2006) Optimal block design for asynchronous wake-up schedules and its applications in multihop wireless networks. *IEEE Trans Mob Comput* 5(9):1228–1241

Zhou Z, Peng Z, Cui J, Jiang Z (2012) Handling triple hidden terminal problems for multichannel MAC in long-delay underwater sensor networks. *IEEE Trans Mob Comput* 11(1):139–154

Asynchronous Wake-Up Coordination Scheme/Protocol

► [Asynchronous Coordination Techniques](#)

Asynchronous Wake-Up Scheme/Protocol

► [Asynchronous Coordination Techniques](#)

Auction

Yanjiao Chen¹ and Qian Zhang²

¹School of Computer Science, Wuhan University, Wuhan, People's Republic of China

²Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, People's Republic of China

Synonyms

[Bidding](#); [Transaction](#)

Definitions

Auction is a concept from microeconomics, referring to the practice of buying and selling goods through the process of bidding. Auction participants include buyers, sellers, and the auctioneer, and the role of the auctioneer may be assumed by

the seller, especially if there is a single seller. An auction mechanism is a series of rules specifying how an auction is conducted. The three most important components of an auction mechanism are the determination of the winners, the allocation of items, and the payment.

Historical Background

Auction is deemed as an effective way to assign items to buyers who value them the most. According to different numbers of participants, auctions can be categorized into three settings.

- *Forward auction*. One seller, who also acts as the auctioneer, and multiple buyers.
- *Reverse auction*. One buyer, who also acts as the auctioneer, and multiple sellers.
- *Double auction*. Multiple buyers, multiple sellers, and a third-party auctioneer.

In the forward auction, if there is a single piece of item, there are four standard auction mechanisms.

- *English auctions* or open ascending-bid auctions. Buyers openly bid against each other by iteratively raising their bids to be higher than the current highest bid. The auction terminates when no buyer offers a higher bid, and the buyer with the highest bid wins the item and pays the corresponding price.
- *Dutch auctions* or open descending-bid auctions. The auctioneer starts with a high bid and then lowers the bid until some buyer accepts and pays the corresponding price.
- *First-price sealed-bid auctions*. Buyers bid simultaneously without knowing each other's bids, i.e., sealed bids. The buyer with the highest bid wins the item and pays the corresponding price. Different from the English auction, buyers can only submit their bids once and cannot adjust the bids according to other buyer's bids.

- *Second-price sealed-bid auctions.* This type of auction is identical to the first-price sealed-bid auctions, with the exception that the winner pays the second-highest bid.

These four types of auctions can be easily extended to reverse auctions and to accommodate multiunit items, and there have been many variants and extensions of these standard auction mechanisms. Besides, there are several other more advanced auction mechanisms.

- *Combinatorial auctions.* Combinatorial auctions allow buyers to bid on combinations of items, such that the bid on a combination may not equal the sum of bids on individual items in the combination. For example, a buyer bids \$1 on item A and \$2 on item B , but may bid \$2.5 on the combination $\{A, B\}$ if A and B are substitutes, or bid \$4 on the combination $\{A, B\}$ if A and B are complements to each other.
- *All-pay auctions.* Normally, nonwinning buyers or sellers will not pay in auctions. All-pay auctions exert a compulsory payment on every participant and may be used to model entry fees or other costs in the process of auctions.
- *VCG auctions* (Vickrey 1961; Clarke 1971; Groves 1973). Vickrey-Clarke-Groves (VCG) auction mechanism tries to maximize social welfare with feasible allocations that satisfy the constraints of the auction (e.g., total number of auctioned items). Take forward auction as an example. The auctioneer finds one optimal feasible allocation A^* that maximizes the total bids of all winning buyers (usually through brute force). The payment of a winning buyer i with bid b_i is calculated as follows. The auctioneer removes buyer i and finds the optimal feasible allocation $\tilde{A}^*$ among the rest of the buyers. Then, buyer i will be charged the price of $\sum_{j \neq i} b_j(\tilde{A}^*) - \sum_{j \neq i} b_j(A^*)$, i.e., the difference of the utility of other buyers when buyer i participates in the auction or not. The VCG mechanism can achieve truthfulness and social welfare maximization, but its major drawback is high computational complexity, since it is often NP-hard to find the optimal feasible allocation.
- *McAfee auctions* (McAfee 1992). The McAfee auction mechanism is developed for single-unit or multiunit double auctions. Take the single-unit auction, for instance. Based on their bids, the auctioneer sorts the sellers in a non-descending order and the buyers in a non-ascending order. The auctioneer finds k , the last profitable pair of seller and buyer, i.e., the bid of the k -th buyer is higher than that of the k -th seller, but the bid of the $(k + 1)$ -th buyer is lower than that of the $(k + 1)$ -th seller. The first $(k - 1)$ sellers win, and each receives the payment of the bid of the k -th seller; the first $(k - 1)$ buyer win and each pays the bid of the k -th buyer.

Since auction participants are selfish and rational individuals who will adopt the optimal bidding strategy to gain the highest utility, game theory can be applied to analyze auctions. An auction is economic robust if the following three properties are satisfied.

- *Individual rationality.* A buyer or a seller is individually rational in the sense that they will not participate in the auction if the obtained utility is less than the utility of no participation. An auction mechanism is individually rational, if all sellers and buyers achieve non-negative utility. This usually means that any seller is paid more than its bid, and any buyer pays less than its bid.
- *Truthfulness or strategy-proofness.* The buyers and sellers have the motivation to manipulate their bids to maximize their own utilities. Being truthful means that a seller or a buyer will submit a bid that equals their true valuation for the item. A truthful auction mechanism guarantees that a buyer or a seller cannot get a higher payoff by misreporting their true valuations; thus they will have no incentive to be untruthful.
- *Budget balance.* Budget balance is often considered in the double auction. It means that the auctioneer maintains nonnegative budget. In

other words, the payment that the auctioneer receives from all buyers is no less than the payment given to all sellers. For regulators, budget balance is often enough to motivate them to host spectrum auctions. The profit-oriented auctioneers, however, may aim at revenue maximization.

Apart from economic robustness, there are several other concerns in the design of auction mechanisms.

- *Social welfare.* Generally speaking, social welfare is the utility of all auction participants, including the buyers, the sellers, and the auctioneer. As the payment merely transfers utility among auction participants, it does not contribute to the social welfare. Therefore, the social welfare can be calculated as the difference between the total true valuation of winning buyers and the total true valuation of winning sellers. In a truthful auction, the social welfare is simply the total bid of winning buyers minus the total bid of winning sellers.
- *Collusions.* Auction participants may collude to rig the auction and gain a higher utility. For example, buyers may form collusions to bid against other buyers but not the buyers in the collusion, resulting in a favorable allocation and a lower payment. The seller may collude with the auctioneer to insert dummy bids to gain a higher payment. The practice of collusion in auctions has been revealed by many empirical studies, calling for effective countermeasures.
- *Privacy-preserving.* The bid of a buyer may be sensitive and private information that should be shielded from the non-trustworthy auctioneer and other rival bidders. For example, in a truthful auction, a buyer's bid is its true valuation, which may be closely related to its profit of winning the spectrum. The disclosure of sensitive information will render unbalanced advantage to the informed entities and cause economic damage to those whose information is divulged. Therefore, privacy-

preserving auction mechanism design is gaining more and more attention.

Key Applications

The most successful application of auctions in wireless networks is dynamic spectrum allocation. Spectrum is an indispensable resource for wireless communications. To address the underutilization problem resulted from static spectrum allocation, it is proposed to dynamically redistribute spectrum through auctions. Different from traditional goods, spectrum features interference-constrained spatial reuse, i.e., the same channel can be reused by multiple non-interfering buyers whose transmission ranges do not overlap. Therefore, the objective of spectrum auctions is to realize spectrum reusability while achieving economic robustness or social welfare maximization. The interference relationship of buyers is usually represented by the interference graph, an undirected graph constructed based on the transmission range of the spectrum and geographic information of the buyers. A typical interference graph can be denoted as $G = (V, E)$, in which the set of nodes V represents the set of buyers and the set of edges E represents interference relationship. If two buyers interfere with each other, there is an edge between them; otherwise, there is no edge between them.

The following are some typical spectrum auction mechanisms.

- *Forward spectrum auction* (Zhou et al. 2008; Jia et al. 2009). The single spectrum owner with M channels acts as the auctioneer, and each buyer can demand more than one channels. After sorting the buyers according to their bids in a non-ascending order, the auctioneer will sequentially check each buyer. If the buyer's demand d_i is fewer than the available channels $M - e_i$, in which e_i is the number of channels allocated to the buyer's interfering neighbors, the buyer will become a winner. A winning buyer will pay according to the critical value, which is the lowest value that the buyer has to bid in order to win.

- *Double spectrum auction* (Zhou and Zheng 2009). Assume that each buyer demands one channel and each seller owns one channel. The third-party auctioneer first divides buyers into groups, each of which only contains non-interfering buyers. The group bid is determined by the group size and the minimum bid in the group. Regarding each buyer group as a virtual buyer, the auctioneer can execute the McAfee auction mechanism to determine the spectrum allocation and the payment.
- *Heterogeneous spectrum auction* (Feng et al. 2012; Chen et al. 2014). The difference between homogeneous and heterogeneous spectrum auctions is that heterogeneous spectrum auction groups non-interfering buyers on a per-channel basis, since the interference relationship of buyers on different channels is different.
- *Combinatorial spectrum auction* (Zheng et al. 2015). Motivated by the fact that channels of contiguous frequency are easier to operate, combinatorial spectrum auction allows buyers to bid on combinations of channels. To realize spatial reuse, the same channel in the combinations of two non-interfering buyers is represented by two different virtual channels. Then, buyers are sorted in the non-ascending order according to the ratio of the bid to the combination size. The auctioneer then sequentially checks each buyer and selects winners as those whose demanded combination does not contain any channel that has already been allocated. A winning buyer will also pay its critical value, as in the forward spectrum auction.
- *Online spectrum auction* (Wang et al. 2010). Online spectrum auctions involve the temporal dynamics of spectrum demand and supply and require buyers to specify their requested time slots. At each time slot, the auctioneer collects bids from the arriving buyers and examines the availability of channels. Given the current buyers, the auctioneer will first conduct a screening process to remove the buyers whose bids are lower than the estimated future value of the channel. Then, the auctioneer can determine spectrum allocation

among the remaining buyers using the double spectrum auction mechanism.

Apart from spectrum distribution, auction has been applied to other scenarios in wireless networks.

- *Power allocation*. Auctions can enable wireless users to cooperate in distributed power allocation to improve the overall network performance (Liu et al. 2013).
- *Crowdsensing*. Mobile crowdsensing platforms can adopt auctions to incentivize the participation and high-quality works from crowd workers (Luo et al. 2017).
- *Mobile data offloading*. Wireless service providers can use auctions to employ Wi-Fi or femtocell access points for data offloading (Iosifidis et al. 2015).
- *Video streaming*. Auctions can be used to motivate nearby wireless users to cooperate in video downloading and sharing, thus improving wireless video streaming performance (Han et al. 2009).

Cross-References

- [Collusion](#)
- [Matching for Cooperative Spectrum Sharing](#)

References

- Chen Y, Zhang J, Wu K, Zhang Q (2014) TAMES: a truthful double auction for multi-demand heterogeneous spectrums. *IEEE Trans Parallel Distrib Syst* 25(11):3012–3024
- Clarke EH (1971) Multipart pricing of public goods. *Public Choice* 11(1):17–33
- Feng X, Chen Y, Zhang J, Zhang Q, Li B (2012) TAHES: a truthful double auction mechanism for heterogeneous spectrums. *IEEE Trans Wirel Commun* 11(11):4038–4047
- Groves T (1973) Incentives in teams. *Econ J Econ Soc* 41(4):617–631
- Han Z, Su GM, Wang H, Ci S, Su W (2009) Auction-based resource allocation for cooperative video transmission protocols over wireless networks. *EURASIP J Adv Sig Process* 2009:4

- Iosifidis G, Gao L, Huang J, Tassiulas L (2015) A double-auction mechanism for mobile data-offloading markets. *IEEE/ACM Trans Netw (TON)* 23(5):1634–1647
- Jia J, Zhang Q, Zhang Q, Liu M (2009) Revenue generation for truthful spectrum auction in dynamic spectrum access. In: *The 10th international symposium on mobile ad hoc networking and computing*. ACM, pp 3–12
- Liu Y, Tao M, Huang J (2013) An auction approach to distributed power allocation for multiuser cooperative networks. *IEEE Trans Wirel Commun* 12(1):237–247
- Luo T, Kanhere SS, Huang J, Das SK, Wu F (2017) Sustainable incentives for mobile crowdsensing: auctions, lotteries, and trust and reputation systems. *IEEE Commun Mag* 55(3):68–74
- McAfee RP (1992) A dominant strategy double auction. *J Econ Theory* 56(2):434–450
- Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. *J Financ* 16(1):8–37
- Wang S, Xu P, Xu X, Tang S, Li X, Liu X (2010) TODA: truthful online double auction for spectrum allocation in wireless networks. In: *IEEE symposium on new frontiers in dynamic spectrum*. IEEE, pp 1–10
- Zheng Z, Wu F, Chen G (2015) A strategy-proof combinatorial heterogeneous channel auction framework in noncooperative wireless networks. *IEEE Trans Mob Comput* 14(6):1123–1137
- Zhou X, Zheng H (2009) TRUST: a general framework for truthful double spectrum auctions. In: *International conference on computer communications*. IEEE, pp 999–1007
- Zhou X, Gandhi S, Suri S, Zheng H (2008) eBay in the sky: strategy-proof wireless spectrum auctions. In: *The 14th annual international conference on mobile computing and networking*. ACM, pp 2–13

Auction Mechanism

- [Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement](#)

Authenticated Query Processing

- [Verifiable Cloud Computing](#)

Authentication

Si Chen

West Chester University, West Chester, PA, USA

Synonyms

[IEEE 802.11](#); [IEEE 802.1x](#); [Wireless credentials](#); [WLAN authentication](#)

Definitions

Authentication methods in wireless networks allow the wireless access point (AP) or broadband wireless router to authorize individuals and devices, and only those who get authorized can access the network. For maximum security, client devices should also authenticate to the wireless network using pre-shared key (PSK) or Extensible Authentication Protocol (EAP) authentication. The use of mutual authentication is essential in a wireless network. This will guard against many security issues, such as man-in-the-middle (MITM) attacks. With authentication, the wireless client and the wireless AP must prove their identity to each other.

Historical Background

When setting up a wireless network, the network administrator has to have some way to ensure that wireless clients, and the devices they are using, are trusted. A typical solution to this question is to use a standard kind of authentication, which in today's wireless network area means IEEE 802.11 and 802.1x authentication.

The first wireless network standard, 802.11, permits a network to be made relatively secure if the user establishes their wireless network using the Wired Equivalent Privacy (WEP) protocol (IEEE 1999). Since the data is transmitted over the air, it is a simple task for a malicious device to sniff it and grab sensitive information. To avoid this type of attack, WEP consists of a

passphrase that uses secret and shared encryption keys that are generated by the wireless host and then passed to the wireless client. These keys are used for encrypting all the data that travels across the airwave and thus thwarting anyone trying to sniff the network. In WEP, there are two authentication mechanisms: (1) open system authentication (OSA) and (2) pre-shared key authentication (PSK). However, it is well known that the IEEE 802.11 WEP protocol is vulnerable to two types of attacks: (1) message privacy and message integrity attacks (Johnson and Maltz 1996) and (2) probabilistic cipher key recovery attacks, such as the Fluhrer-Mantin-Shamir attack (Fluhrer et al. 2001; Stubblefield et al. 2002). In 2003, the Wi-Fi Alliance announced that WEP had been superseded by Wi-Fi Protected Access (WPA) protocol. WPA protocol has the addition of the Temporal Key Integrity Protocol (TKIP). TKIP avoids the problems of WEP by implementing per-packet key mixing with a rekeying system and message integrity check (MIC) mechanism. In 2004, as an updated version of WPA, WPA2 was released along with the IEEE 802.11i standard. Both WPA and WPA2 support the PSK authentication key distribution mechanism and use a stronger encryption than WEP. It is designed for home and small office networks and doesn't require an authentication server.

Another wireless network standard, the IEEE 802.1x, is widely adopted for large wireless area networks such as a hotel or big company. It takes the predecessor IEEE 802.11 one step further and includes computer and user identification, dynamic key creation, and centralized authentication. IEEE 802.1x supports the Internet Authentication Service (IAS) which uses the Remote Authentication Dial-In Service (RADIUS) protocol. In addition, 802.1x ties to the Extensible Authentication Protocol (EAP) which supports multiple authentication methods such as Kerberos, public key authentication, and certificates.

Many security issues have been found in these wireless authentication protocols. In 2008, an attack on TKIP (Beck and Tews 2009) was released which allows an adversary exploiting WPA's MIC. In 2017, security researchers have discovered a significant vulnerability, named

KRACKs (Vanhoeft and Piessens 2017), in WPA2. It uses a group of vulnerabilities that exist in the implementation of the authentication protocol to launch key reinstallation attack and eventually allows the attacker to intercept and steal data. In January 2018, Wi-Fi Alliance announced the release of WPA3 with several security improvements over WPA2.

Foundations

There are three main methods of authentication that are used on wireless networks.

- **Open System Authentication:** Open system authentication allows any wireless device to connect to the network as long as the end device knows the service set identifier (SSID) used on the network. The significant advantage of this type of authentication is its simplicity: any client can directly connect without complicated configuration. The problem with this authentication scheme is that the SSID is typically broadcast so any device within the wireless coverage range can join the network. Moreover, the SSID can be obtained from the association request frames even if the access point is set not to broadcast the SSID.
- **Pre-Shared Key Authentication:** The pre-shared key authentication method is commonly used on individual and small business wireless networks. It uses a pre-shared key that is distributed to the client device before establishing the connection. The PSK allows any wireless client who has the key to join the network. The client also has the option to use the key to encrypt the data. This can protect sensitive data traveling across the wireless network from becoming captured. The original 802.11 WEP protocol utilizes RC4 for encryption and has been deprecated due to several vulnerabilities. The improved WPA protocol ties with the TKIP which utilizes dynamic keys that were not supported with WEP for encryption. However, a similar vulnerability has been found since TKIP uses the same procedures that WEP does. WPA2

replaced TKIP with Counter Mode Cipher Block Chaining Message Authentication Code Protocol (CCMP) which is primarily based on Advanced Encryption Standard (AES), and it is proven to provide strong encryption.

A PSK is relatively easy to configure though it requires some configuration at the client side. However, a wireless network based on a PSK does not scale well. With a large number of clients, it becomes more difficult to periodically update the PSK.

- **EAP (Extensible Authentication Protocol) Authentication:** EAP authentication is an arbitrary authentication mechanism that is used in the wireless network, and it is part of the 802.1x standard. EAP allows for several choices of authentication methods including MD5, EAP-TLS, EAP-TTLS, and PEAP. In order to use EAP to authenticate users when they connect to the wireless network, a third-party authentication server is required at time of the client association, such as RADIUS. The wireless host opens the client's port for other types of traffic based on access rights stored in the authentication server.

Key Applications

Authentication is one of the most important parts of wireless protocol which is used for establishing user identities and for securing the wireless communication.

Cross-References

- [Access Control](#)
- [Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs](#)
- [Blockchain: Enabling Future Internet with Intrinsic Security](#)

References

Fluhrer S, Mantin I, Shamir A (2001) Weaknesses in the key scheduling algorithm of rc4. In: International workshop on selected areas in cryptography. Springer, Berlin, Heidelberg, pp 1–24

IEEE (1999) Wireless lan medium access control (mac) and physical layer (phy) specifications. IEEE Std 802.11

Johnson DB, Maltz DA (1996) Dynamic source routing in ad hoc wireless networks. In: Mobile computing. Springer, Boston, MA, pp 153–181

Stubblefield A, Ioannidis J, Rubin AD (2002, February) Using the Fluhrer, Mantin, and Shamir Attack to Break WEP. In: Proceedings of the Network and Distributed System Security Symposium (NDSS), San Diego, California, USA

Tews E, Beck M (2009) Practical attacks against WEP and WPA. In: Proceedings of the second ACM conference on Wireless network security (WiSec '09). ACM, New York, NY, USA, 79–86. <https://doi.org/10.1145/1514274.1514286>

Vanhoef M, Piessens F (2017) Key reinstallation attacks: Forcing nonce reuse in wpa2. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. ACM, Dallas, Texas, USA, pp 1313–1328

Authentication in Delay-/Disruption-Tolerant Networks

- [Security in Delay-Tolerant Networks](#)

Automotive Cybersecurity

- [Automotive Security](#)

Automotive Security

Nevrus Kaja, Ahmad Nasser, Di Ma, and Adnan Shaout
University of Michigan, Dearborn, MI, USA

Synonyms

[Automotive cybersecurity](#); [In-vehicle security](#); [Vehicle cybersecurity](#)

Definitions

Automotive Security is the field of assessing cyber-threats against automotive systems and developing security countermeasures for detecting, preventing the corresponding attacks, and reacting to a security breach. As modern vehicles become connected, the threat of attack becomes increasingly real. Coupled with advances in driver assistance and autonomous solutions, the impact on security is further expanded. Thus, automotive security aims at ensuring automotive systems can achieve their intended function without the danger of being maliciously manipulated by an attacker. A few definitions used in vehicle security are provided below:

- **Cyber-physical system:** A system of collaborating computational elements controlling physical entities
- **Cybersecurity:** An attribute of a cyber-physical system that relates to avoiding unreasonable risk due to an attack (Committee SVESS et al. 2016)
- **Attack:** Exploitation of vulnerabilities to obtain unauthorized access to or control of assets with the intent to cause harm
- **Threat:** A circumstance or event with potential to cause harm
- **Exploit:** An action who uses vulnerability to cause undesired, unexpected, or unanticipated behavior
- **CAN:** Control Area Network is a dominant communication network protocol used for intra-vehicle communication

Historical Background

Automotive security related threats initially came to light with the introduction of keyless entry systems. For such vehicles, the primary threat was that of theft, and vehicle manufacturers were keen on implementing security features to prevent a thief from wirelessly unlocking the vehicle and driving off (Biham et al. 2007). Security features included rolling codes as well

as immobilizer units that can prevent the vehicle from starting unless the correct key is present. Another area that initially gained recognition was the tire pressure monitoring sensors (TPMS) (Heary 2010). These sensors gave off RF signals that could be used to allow tracking of vehicles, which would violate the driver's privacy. As vehicles added more connectivity features through a telematics unit, USB devices, bluetooth, and Wi-Fi, the attack vectors increased drastically and so did the interest of the research community to uncover potential exploits that could be found in these systems. The work of Checkoway et al. (2011) was instrumental in showing the extent to which attacks could be launched against vehicles through various attack surfaces like OBD tools, CD Roms, and more. This brought cybersecurity to the forefront in the minds of vehicle OEMs who knew that security needed more attention. The attacks by Miller and Valasek (2015) launched remotely on the Jeep in 2015 further demonstrated that the threats were not only academic but real, and lead FCA to ship USB sticks with patches to close the security vulnerabilities uncovered by the attacks. Even before that date, several standardization bodies were aware of the need to address cyber threats on vehicle systems. One of the early ones was EVITA, a European Commission tasked with defining security use cases and deriving security requirements for vehicle systems. EVITA produced 18 use cases and defined a security architecture for on-board automotive networks as well as a threat assessment method.

In addition to EVITA, several other industry initiatives were created to address automotive security. Some of these are provided below:

- **SAE J3061:** The first cyber-security framework to address all the main security aspects of producing secure automotive systems (Committee SVESS et al. 2016).
- **C2C:** Car to Car communication consortium aims to create a European industrial standard for communicating cars.
- **ETSI C-ITS:** A not-for-profit organization aims to achieve standards for Cooperative Intelligent Transport Systems.

- USDOT/CAMP: A joint initiative to produce a security credential management system (SCMS).
- AUTOSAR WP-X: The security working group within AUTOSAR produces software specifications to define security functions in embedded systems.
- UPTANE: An academic and industry co-initiative to standardize Over-the-Air updates in vehicles.
- AUTO-ISAC: An industry and government co-initiative for information sharing across the automotive supply chain relating to cyber threats (Auto ISAC 2017).
- ISO and SAE joint working group to produce a common cyber security standard: ISO-SAE 21434.
- IEEE Standard for Wireless Access in Vehicular Environments (IEEE 1609.2).

Foundations

To study automotive security, one usually starts with threat analysis and risk assessment to identify attack surfaces, attack agents, security objectives, and controls.

Threat Analysis and Risk Assessment processes: In a fashion that parallels the hazard and risk assessment of functional safety (ISO 2011), automotive security requires the enumeration of threats and the qualification of the corresponding risk. There exist several threat assessment methods such as EVITA-THROP (Henniger et al. 2009), OCTAVE (Alberts et al. 2003), TVRA (IMS 2008), HEAVENS (Islam et al. 2014), and NHTSA composite modeling (McCarthy et al. 2014). Once the threats are identified and the corresponding risk is calculated, product developers derive security requirements to address the associated threats. The higher the risk of an attack and its impact, the more stringent the security level would be, which means added hardware countermeasures, and a stricter security profile.

Attack Surfaces: With added connectivity comes the risk of added exposure to security threats due to the ever expanding attack surface. As shown in Fig. 1, today's vehicles offer

numerous attack surfaces. In order to address these cyberthreats, automotive security engineers create security defenses both at the vehicle boundary and at the internal vehicle layers. Figure 1 provides an overview of vehicle communication interfaces. Dotted lines represent wireless connections, while solid lines represent wired connections.

Attack Agents: An important element in automotive security is the agent who is performing the attack. Different agents will have different capabilities, resources, expertise and motivations. For example, a mechanic has far less capabilities and different attack motivations from a nation state hacker. The following is a list of different attack agents who might compromise a vehicle.

Threat Agent

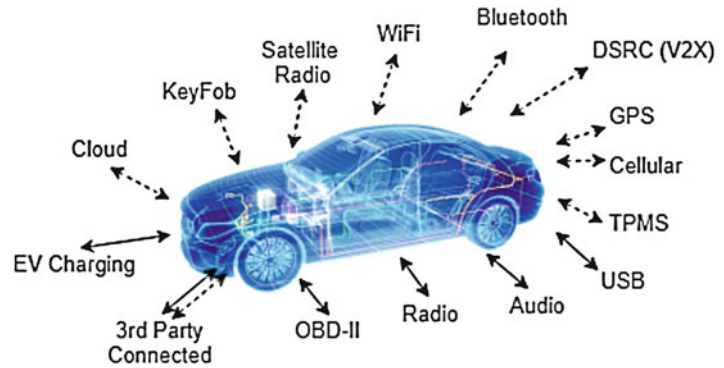
- Thief
- Vehicle Owner
- Mechanic
- Hacktivist
- Organized Crime
- Researcher
- Terrorist
- State sponsored

Security Objectives: Automotive security is different from IT security due to the dimension of control which results in security threats impacting safety. Still, several security objectives are typical of IT systems with the exception that the assets being protected are automotive specific. Among the primary security objectives for automotive systems are:

- Prevent attacks that have an impact on vehicle safety
- Protect driver privacy and prevent vehicle tracking
- Protect vehicle OEM reputation and brand
- Protect algorithms and IP especially ones related to Autonomous and ADAS systems, from theft through reverse engineering

Throughout the automotive industry, there is a general consensus that automotive security

Automotive Security,
Fig. 1 Attack surfaces



A

requires a security-in-depth approach. This translates to providing security at each layer of a vehicle interface (known as the layered approach). The first layer being the vehicle external interface (telematics unit, OBD, infotainment, etc.). The second layer being the vehicle network gateway that separates different vehicle domains and allows for filtering network data inside the vehicle. The third layer being the vehicle network within each domain. And the fourth layer being the Electronic Control Unit itself, primarily the microcontroller unit, sensors, and the software that goes along with it. Building defenses at each one of these layers is meant to increase the attack difficulty to a point that discourages an attacker from attempting to compromise a vehicle system, as the payoff would not justify the required effort.

Security Primitives and Key Applications

To ensure automotive security, researchers in the industry and academia have come up with a range of security requirements that apply to automotive ECUs:

- **Secure Debug:** Accessing to JTAG as well as any debug interface shall be locked to prevent reverse engineering and theft of vehicle IP.
- **Secure Diagnostics:** Vehicles contain extensive diagnostic services meant to allow the OEM to fully diagnose an ECU failure to aid in vehicle maintenance as well as test for

emissions. Such diagnostics shall be protected against attackers who can use these services to manipulate the vehicle.

- **Secure Programming:** Automotive ECUs rely on flash bootloaders to update vehicle software. All application and calibration data shall be cryptographically signed to prevent an attacker from modifying the vehicle software.
- **Secure Communication:** Automotive systems are highly reliant on CAN as the primary network protocol. Since CAN has no built-in security features, it is important to define security protocols that can prevent an attacker from spoofing control or status messages on the CAN bus.
- **Secure Boot:** Checking the authenticity of vehicle software after each boot cycle provides added assurance that vehicle software was not tampered with while at rest.

The above security requirements apply to a variety of automotive ECUs. The below list highlights specific security application areas:

- **Telematics:** Provide various forms of connectivity such as cellular and Wi-Fi. Therefore, they open the door to remote-based attacks. Such systems require strong firewalls to block unwanted traffic into the vehicle.
- **Infotainment:** Such systems are primarily used to provide audio/video and navigation to the end user. By their nature, they use rich operating systems like Android and Linux. To isolate such applications from critical vehicle systems, sandboxing and trustzones are important

to provide isolation between the trusted and untrusted world.

- V2X: Vehicle-to-Everything is another emerging set of communication protocols. Such systems have special requirements for processing a large number of ECC calculations to authenticate other entities. Due to the high potential impact to safety, the hardware in such systems is expected to support FIPS 140-2 level 3 or higher HSMs.
- Vehicle Gateway: Network domains separate vehicle functions into separate CAN bus systems. In this scenario, methodologies such as network isolation and intrusion detection systems are used to detect and prevent outside and inside intruders (Kaja et al. 2017; Zhang et al. 2018).
- In-Vehicle Network: In-Vehicle networks enable ECU-to-ECU communication via robust and reliable networks. In order to secure such networks, different mechanisms are used such as AUTOSAR SecOC, for authentication, and Freshness.
- Sensors: The number of sensors in a vehicle has ballooned to allow understanding of its surrounding. Authenticating sensor data and source is important to prevent sensor spoofing or sensor counterfeiting.
- AUTOSAR based ECUs: As one of the most widely prevalent automotive software platform, AUTOSAR now presents a rich area of study in automotive embedded security. Therefore, applying best practices within AUTOSAR for increasing the security resilience of such systems is mandatory (Nasser et al. 2017)

Cross-References

- [Dedicated Short-Range Communication \(DSRC\)](#)
- [Industrial Cyber-physical Systems](#)
- [Internet of Things: Architecture, Key Applications, and Security Impacts](#)

- [Network Security](#)
- [Vehicular Networks](#)
- [Wireless Data Security](#)

References

- Alberts C, Dorofee A, Stevens J, Woody C (2003) Introduction to the octave approach. Technical report. Carnegie-Mellon University
- Auto ISAC (2017) Automotive information sharing and analysis center. <https://www.automotiveisac.com/>. Accessed 6 May 2018
- Biham E, Dunkelman O, Indesteege S, Keller N and Preneel B (2007) How to steal cars – A practical attack on keeLoq. In: *Crypto 2007*, Santa Barbara, California
- Checkoway S, McCoy D, Kantor B, Anderson D, Shacham H, Savage S, Koscher K, Czeskis A, Roesner F, Kohno T et al (2011) Comprehensive experimental analyses of automotive attack surfaces. In: *USENIX security symposium*, San Francisco
- Committee SVESS et al (2016) SAE j3061-cybersecurity guidebook for cyber-physical automotive systems. SAE-Society of Automotive Engineers
- Heary J (2010) Defcon: Hacking tire pressure monitors remotely. <https://www.networkworld.com/article/2231495/cisco-subnet/defcon—hacking-tire-pressure-monitors-remotely.html>
- Henniger O, Ruddle A, Seudié H, Weyl B, Wolf M, Wollinger T (2009) Securing vehicular onboard IT systems: the EVITA project. In: *VDI/VW automotive security conference*, Ingolstadt
- IMS (2008) Final draft etsi es 282 007 v2. 0.0 (2008-03)
- Islam M, Sandberg C, Bokesand A, Olovsson T, Broberg H, Kleberger P, Lautenbach A, Hansson A, Söderberg-Rivkin A, Kadhivelan S (2014) Deliverable d2-security models. HEAVENS Project, Version 1
- ISO (2011) 26262: road vehicles-functional safety. International Standard ISO/FDIS 26262
- Kaja N, Shaout A, Ma D (2017) A two stage intrusion detection intelligent system. In: *The International Arab Conference on Information Technology*, Tunisia
- McCarthy C, Harnett K, Carter A (2014) Characterization of potential security threats in modern automobiles: a composite modeling approach. Technical report
- Miller C, Valasek C (2015) Remote exploitation of an unaltered passenger vehicle. *Black Hat USA 2015*
- Nasser AM, Ma D, Muralidharan P (2017) An approach for building security resilience in AUTOSAR based safety critical systems. *J Cyber Secur Mobil* 6(3):271–304
- Zhang L, Shi L, Kaja N, Ma D (2018) A two-stage deep learning approach for CAN intrusion detection. *NDIA ground vehicle systems engineering and technology symposium*, Michigan

Bandwidth Allocation in Data Center Networks

Haojun Yang and Kan Zheng
Intelligent Computing and Communications (IC²) Lab, Wireless Signal Processing and Networks (WSPN) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Synonyms

[Resource allocation](#)

Definition

In data center networks (DCNs), infrastructure as a service (IaaS) providers can flexibly allocate bandwidth in order to achieve diverse benefits for providers and tenants. For the IaaS providers, appropriate schemes of bandwidth allocation can improve the utilization of DCNs, and thus more revenues can be generated. On the other hand, the amount of bandwidth allocated for the tenants largely determines the performance of applications and the fairness of coexisting tenants.

Historical Background

In recent years, cloud computing is widely developed for providing various Internet services such as search, video caching, and distributed storage. To efficiently meet the requirements of data-intensive cloud computing applications, large-scale DCNs are built by IaaS providers. Among diverse studies of DCNs, bandwidth allocation plays a key role; it has been investigated from different perspectives, achieving many advances (Bari et al. 2013; Del Piccolo et al. 2016; Xia et al. 2017; Zhang et al. 2018).

As the most common and simple scheme, the best effort manner of bandwidth allocation is first introduced. In general, the best effort allocation cannot satisfy the explicit demands specified by tenants. As a result, malicious tenants can unfairly improve the network performance by establishing multiple Transmission Control Protocol (TCP) connections or simply using User Datagram Protocol (UDP) (Xia et al. 2017), which leads to the bandwidth guarantee of normal tenants not being ensured. To this end, fairness is the future research focus in the best effort sharing.

In contrast to the best effort sharing, static bandwidth reservation scheme can provide bandwidth guarantee for each tenant according to its requirement (Chen et al. 2014) and thus is

widely studied recently. The obvious advantage of deterministic allocation is that the bandwidth available to a tenant is irrespective of the network traffic from other tenants. However, when tenants do not fully make use of their bandwidth, static bandwidth reservation scheme will lead to the low utilization of DCNs. Therefore, improving the utilization is crucial issue in bandwidth guarantee-like schemes.

Foundations

Requirements of Bandwidth Allocation

Bandwidth allocation in the DCNs generally depends on many aspects including guarantee, fairness, and utilization (Sun et al. 2016; Chen et al. 2014).

Guarantee is an important requirement affecting the performance of DCNs. The static scheme mentioned above can provide perfect performance isolation among all the tenants. Therefore, it is also referred to as full bandwidth guarantee. In order to further improve the utilization of full bandwidth scheme, a flexible method is minimum bandwidth guarantee, which enables tenants to achieve worst-case performance and thus to bound costs.

When guarantee cannot be ensured, fairness will be the crucial requirement in bandwidth allocation. In public clouds, fairness allocation means that bandwidth is divided in proportion to the payments among all tenants. While in private clouds, the abovementioned definition of fairness is no longer applicable. Specifically, fairness generally depends on the performance of different applications, i.e., the priority and quality of service should be considered.

It is vital to fully utilize the available bandwidth in DCNs. Intuitively, achieving high utilization is to find the optimal bandwidth allocation by various traffic control methods. The idle bandwidth should be used for meeting application demands as far as possible in private clouds. With regard to public clouds, IaaS providers should establish encouragement

mechanisms for tenants to improve the utilization of bandwidth.

Solutions of Bandwidth Allocation

Minimum Guarantee Bandwidth Allocation

Minimum bandwidth guarantee can provide the bandwidth specified in a service-level agreement, achieving the worst-case performance. Two representative static bandwidth reservation solutions are first introduced, i.e., SecondNet and Oktopus can ensure per-flow and per-tenant bandwidth guarantee, respectively (Guo et al. 2010; Ballani et al. 2011).

However, maintaining a fixed bandwidth is not always practical in the real world. Therefore, based on the time-varying requirements, Proteus proposes a virtual network abstraction model called time-interleaved virtual clusters (Xie et al. 2012). The bandwidth demand on each virtual link is described as a time-varying function. Compared with the static reservation solutions, Proteus can significantly improve the utilization of DCNs and reduce the costs on bandwidth.

Jeyakumar et al. put forward the EyeQ solution which provides predictable minimum bandwidth by rate control at senders according to the feedback from receivers (Jeyakumar et al. 2012, 2013). Moreover, Jeyakumar et al. build a test bed to evaluate the EyeQ performance. Another feature of EyeQ is that the rate control is distributed executing at the network edge. Hence, its implementation complexity is relatively low.

ElasticSwitch, also as a distributed solution, ensures bandwidth guarantee under the hose model (Popa et al. 2013). There is two-layer allocation in the ElasticSwitch. First of all, the higher layer transforms the guarantees of virtual machines (VMs) into a set of minimum guarantees for VM pairs. In order to enhance the utilization, the lower layer then dynamically shares the spare bandwidth on a link in proportion to the bandwidth guarantees of VM pairs on the link.

Proportional Bandwidth Allocation

When there are not additional service-level agreements of bandwidth, the bandwidth will be allocated in proportion to what a tenant has paid for other hardware resources, ensuring the allocation fairness in DCNs.

In the NetShare solution, each tenant is assigned with a weight, and the bandwidth is allocated based on the weighted proportionality of all tenants (Lam et al. 2010). The paradigm of the tenant-level fairness can be shifted into the levels of any traffic sources confined to a single tenant, including a VM or a process. Hence, Seawall is proposed in Shieh et al. (2011), where each traffic source is also assigned with a network weight. According to the weights, different traffic sources can acquire the bandwidth proportionally. Furthermore, Seawall adopts a TCP-like congestion control mechanism to limit the transmission rate to obtain per-source fairness.

Another solution is FairCloud proposed in Popa et al. (2012), where the bandwidth allocation is based on the number and weights of VMs owned by a tenant. In general, there are three aspects on bandwidth proportionality, namely, proportional sharing at link-level (PS-L), proportional sharing at network-level (PS-N), and proportional sharing on proximate links (PS-P). In the PS-L strategy, the weight of a tenant on a link is calculated as the sum of the weights of this tenant's VMs which communicate at this link. Similarly with the PS-L, the PS-N policy mainly focuses on the weights of source-destination VM pairs, while the PS-P strategy is generally applied in the tree-based network topology to offer per-source fairness. At last, it is noted that FairCloud is not suitable for deploying in a relatively large scale network, since switches require per-VM queue (Chen et al. 2014).

Key Applications

In various data-intensive cloud computing applications supported by DCNs, bandwidth

allocation of tenants and VMs is the operation fundamental for IaaS providers.

Future Directions

Besides the aforementioned three requirements of bandwidth allocation, we discuss two future research directions including, but not limited to, the following.

1. Timeliness-oriented bandwidth allocation: With the emergence of Internet of things (IoT), diverse latency-sensitive applications are gradually being supported by cloud computing, such as automation control of industrial manufacturing or production processes, autonomous driving and remote medical surgery, etc. Unlike the traditional applications, these types of applications pay more attention to the network performance of latency. Therefore, how to conceive low-latency solutions of bandwidth allocation remains as an open problem for larger-scale DCNs.
2. Energy conservation-oriented bandwidth allocation (Ge et al. 2013): As the scale of future cloud computing is continually expanded, the power consumption of large-scale server rooms will be a thorny issue. Hence, studying efficient energy conservation-oriented bandwidth allocation for DCNs is also one of the future research directions; it can significantly reduce environmental pollution.

References

- Ballani H, Costa P, Karagiannis T, Rowstron A (2011) Towards predictable datacenter networks. In: Proceedings of ACM SIGCOMM Conference, Toronto, pp 242–253
- Bari MF, Boutaba R, Esteves R, Granville LZ, Podlesny M, Rabbani MG, Zhang Q, Zhani MF (2013) Data center network virtualization: a survey. *IEEE Commun Surv Tutor* 15(2):909–928

- Chen L, Li B, Li B (2014) Allocating bandwidth in datacenter networks: a survey. *J Comput Sci Technol* 29(5):910–917
- Del Piccolo V, Amamou A, Haddadou K, Pujolle G (2016) A survey of network isolation solutions for multi-tenant data centers. *IEEE Commun Surv Tutor* 18(4): 2787–2821
- Ge C, Sun Z, Wang N (2013) A survey of power-saving techniques on data centers and content delivery networks. *IEEE Commun Surv Tutor* 15(3):1334–1354
- Guo C, Lu G, Yang S, Kong C, Sun P, Wu W, Zhang Y, Wang H, Lv G (2010) SecondNet: a data center network virtualization architecture with bandwidth guarantees. In: *Proceedings of International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, Philadelphia, pp 1–12
- Jeyakumar V, Alizadeh M, Mazières D, Prabhakar B, Kim C (2012) EyeQ: practical network performance isolation for the multi-tenant cloud. In: *Proceedings of USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, Boston, pp 1–6
- Jeyakumar V, Alizadeh M, Mazières D, Prabhakar B, Greenberg A, Kim C (2013) EyeQ: practical network performance isolation at the edge. In: *Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Lombard, pp 297–312
- Lam T, Radhakrishnan S, Vahdat A, Varghese G (2010) NetShare: virtualizing data center networks across services. Technical Report CS2010-0957, Department of Computer Science and Engineering, University of California at San Diego
- Popa L, Kumar G, Chowdhury M, Krishnamurthy A, Ratnasamy S, Stoica I (2012) FairCloud: sharing the network in cloud computing. In: *Proceedings of ACM SIGCOMM Conference*, Helsinki, pp 187–198
- Popa L, Yalagandula P, Banerjee S, Mogul JC, Turner Y, Santos JR (2013) ElasticSwitch: practical work-conserving bandwidth guarantees for cloud computing. In: *Proceedings ACM SIGCOMM Conference*, Hong Kong, pp 351–362
- Shieh A, Kandula S, Greenberg A, Kim C, Saha B (2011) Sharing the data center network. In: *Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Boston, pp 309–322
- Sun X, Ansari N, Wang R (2016) Optimizing resource utilization of a data center. *IEEE Commun Surv Tutor* 18(4):2822–2846
- Xia W, Zhao P, Wen Y, Xie H (2017) A survey on data center networking (DCN): infrastructure and operations. *IEEE Commun Surv Tutor* 19(1): 640–656
- Xie D, Ding N, Hu YC, Kompella R (2012) The only constant is change: incorporating time-varying network reservations in data centers. In: *Proceedings of ACM SIGCOMM Conference*, Helsinki, pp 199–210
- Zhang J, Yu FR, Wang S, Huang T, Liu Z, Liu Y (2018) Load balancing in data center networks: a survey. *IEEE Commun Surv Tutor* 20(3):2324–2352

Base Station Cooperations Under Imperfect Conditions

Jie Gong

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

Synonyms

Coordinated Multipoint (CoMP); Joint precoding/processing; Network MIMO; Virtual MIMO

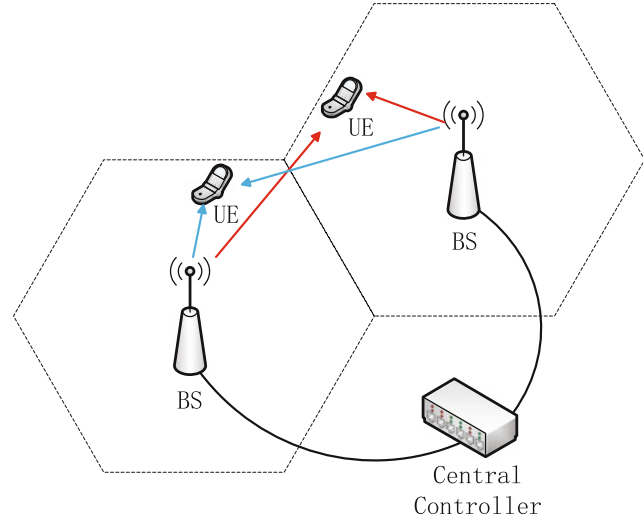
Definitions

Base station (BS) cooperation is a technique by which multiple BSs simultaneously serve multiple user equipments (UEs) in the same frequency band using networked multiple-input-multiple-output (MIMO) protocol. Figure 1 illustrates a simple example of BS cooperation with two BSs and two UEs. The BSs exchange channel state information (CSI) and UEs' data for cooperation via a central controller. Each BS conveys the desired signals of both UEs as shown by the red and blue arrows, respectively. Different from the conventional MIMO technique, BS cooperation is subject to per-BS power constraints instead of a sum power constraint. Suppose M single-antenna BSs cooperatively serve N single-antenna UEs in the downlink, the system model can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \quad (1)$$

where $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$, y_i is the signal received by the i -th UE, $\mathbf{H} \in \mathcal{C}^{N \times M}$ is the integrated channel matrix between BSs and UEs, $\mathbf{x} = [x_1, x_2, \dots, x_M]^T$, x_i is the signal transmitted by the i -th BS under per-BS power constraint $E[x_i] \leq P_i$, and $\mathbf{n} \in \mathcal{C}^N$ is the noise. In BS cooperation, the signal from each BS can contain all the UEs' data. Hence, the transmitted signal can be expressed as $\mathbf{x} = \mathbf{A}\mathbf{d}$, where $\mathbf{A} \in \mathcal{C}^{M \times N}$ is the joint precoding matrix and $\mathbf{d} = [d_1, d_2, \dots, d_N]^T$, d_i is the desired

**Base Station
Cooperations Under
Imperfect Conditions,
Fig. 1** BS cooperation
with two BSs and two UEs



B

data of the i -th UE. One of the key issues in BS cooperation is to design the precoding matrix $\mathbf{A}$ under per-BS power constraint $E[x_i] \leq P_i$.

Historical Background

BS cooperation is dedicated to deal with the inter-cell co-channel interference in 4G/LTE networks. Traditional ways of co-channel interference mitigation either require extra bandwidth for large reuse factor or rely on highly complex signal processing which is not practical for mobile devices. With the development of MIMO technologies and the enhanced BS processing capability, BS cooperation has been proposed to deal with the inter-cell co-channel interference. The basic idea is to treat a group of BSs as a “super-BS” with multiple spatially distributed antennas, also known as network MIMO (Karakayali et al. 2006). By sharing UEs’ CSIs and data, multiple BSs coherently coordinate the transmission and reception; thus other-cell signals are used to assist transmission instead of acting as interference. As a consequence, the network capacity can be greatly enhanced. The idea can be originated to distributed antenna systems (Saleh et al. 1987; Zhang and Andrews 2008). Based on this technique, people have proposed a novel architecture called distributed wireless communication sys-

tem (DWCS) (Zhou et al. 2003), where the antennas are distributively deployed and the baseband processors are jointly managed.

Although BS cooperation is expected to significantly improve network capacity, it encounters enormous challenges in practice when considering imperfect conditions. The major categories of the imperfect conditions are summarized as follows:

- *Limited capacity of central controller.* With the increase on the number of cooperating BSs, a central controller needs to handle the joint precoding for the increased number of UEs. To support high-speed data transmissions, high data processing capacity of the central controller is urgently required.
- *Synchronization.* To realize BS cooperation, symbol-level synchronization among BSs should be guaranteed. However, as the BSs are geographically separated, it is very difficult to achieve high-accuracy synchronization. This difficulty may become even more severe when the number of cooperating BSs increases since the variance of the distances to UEs becomes larger.
- *Limited CSI acquisition capability.* The BS cooperation highly depends on UEs’ CSIs for precoding. Therefore, each BS needs to obtain the CSI of all the UEs, which requires a large amount of resources for channel train-

ing and feedback. However, as the feedback channel capacity is limited, the cooperation performance would degrade due to inaccurate channel feedback.

- *Limited backhaul capacity.* The UEs' data needs to be shared among cooperating BSs, which exponentially increases the burden of backhaul link. BS cooperation cannot be fully realized unless there is sufficient backhaul capacity to support UEs' data sharing.

To achieve the expected capacity gain of BS cooperation in real systems, the above imperfect conditions should be taken into consideration. In the next section, candidate solutions are listed with a summary of the state-of-the-art research progress.

Dealing with Imperfect Conditions

The imperfect conditions involved in the real systems induce novel design in cooperative transmission. In particular, the research efforts are classified into the following three parts.

Clustering with Limited Cooperation Size

Due to the limited capacity of the central controller as well as the difficulty in synchronization, it is impossible to form a "super-BS" over the whole network. On the other hand, it is also unnecessary as the cooperation gain with large inter-BS distance is marginal due to the pathloss effect. Consequently, grouping BSs into clusters is a practical and efficient solution. The clustering algorithms can be classified into two categories. The first category is static clustering (Huang et al. 2009), in which the BS cooperation cluster is pre-determined during network planning stage so that it is easy to implement during network operation stage. However, the problem of static clustering is that the cluster-edge UE is severely degraded due to inter-cluster interference. A modified protocol was proposed in Zhang et al. (2009) by suppressing the interference via inter-cluster cooperation. However, the BS cooperation gain will be sacrificed. Thus, people turn to the second

category of clustering, i.e., dynamic clustering according to UEs' channel conditions. A centralized dynamic clustering algorithm was proposed by Papadogiannis et al. (2008b). It is shown that the dynamic clustering with cluster size 2 can perform the same as the static clustering with size 7. Nevertheless, the centralized algorithm requires a large amount of CSIs as well as high computational capacity of the central controller. An improved algorithm was proposed by Liu and Wang (2009) to reduce the cost on channel acquisition. But still, the limited capacity of the central controller is a performance bottleneck.

To further reduce the computational requirement on the central controller, we proposed a distributed dynamic clustering algorithm (Zhou et al. 2009) where the BSs negotiate with each other distributively to form the clusters. The BSs firstly compute the CSI-related weights to represent the potential gain of joining a certain cluster and then exchange the weights to reach an agreement. It is shown that with a small amount of information exchange among the BSs, the proposed distributed algorithm achieves the same sum-rate performance as the centralized algorithm, but the computational complexity and signaling overhead can be greatly reduced. The dynamic clustering algorithm was further extended to overlapping cluster case (Gong et al. 2011), where a single BS can join multiple clusters simultaneously, and hence, the user fairness can be improved compared with non-overlapping clustering methods.

With clustering technique, the practical constraints including limited capacity of the central controller and synchronization issue can be well solved. Firstly, distributed dynamic clustering reduces the capacity requirement on the central controller. The channel training can be launched locally in each BS, the information exchange can be done by the backhaul link, and the clustering decision is made distributively. Secondly, limited cooperation size reduces the difficulty of synchronization. Simulation results show that the dynamic clustering with sizes 2 to 3 can achieve very high sum data rate. The performance improvement by further increasing the cluster size is marginal. Hence in practice,

clustering with sizes 2 to 3 is widely used. In this case, the synchronization is not a difficult task.

Dynamic CSI Acquisition with Limited Feedback

As each BS requires CSIs from all UEs in the same cluster, CSI acquisition should be redesigned for BS cooperation. In TDD mode, by exploiting the uplink-downlink reciprocity, the CSIs of all UEs can be trained via uplink pilot signal. However, the total training slot length is limited. When splitting the training sequence to the UEs, each of them can only be allocated with limited resources and hence has imperfect CSI. In FDD mode, the CSIs can be trained by each UE via broadcasting a training sequence from a BS and then can be fed back to the BS. In this case, each UE needs to feed back the CSIs of multiple cooperating BSs. Due to limited feedback channel capacity, the feedback channel resource allocation should be carefully designed. In the rest of this subsection, we focus on limited feedback in FDD mode.

Limited feedback for a single link or a single cell has been well studied by Love et al. (2008) and the references therein. Different from single-cell case, however, in BS cooperation scenario, there are different pathloss effects between a specific UE and cooperating BSs. Intuitively, the BSs closer to the UE are more valuable for cooperation. As a result, one can ignore CSI from the far BSs and allocate more feedback bits to the near BSs to enhance the overall performance. Papadogiannis et al. (2008a) proposed a threshold-based CSI feedback algorithm to only feed back the CSIs larger than a certain threshold. However, there lack theoretical results on how to determine the threshold. A more dedicated and realistic model is to use quantization and feed back the CSIs by a given number of bits. Quantized CSI feedback was studied by Zhang and Andrews (2010) and Bhagavatula and Heath (2011) for MIMO interference channel, i.e., the CSIs of other UEs are only used for interference suppression. For cooperation, the performance is more sensitive to the CSIs as the precoding matrix strongly depends on the quantized feedback. Based on the above intuition, more bits

should be allocated to strong links, while fewer bits should be used on weak links, so that the influence of feedback error is minimized. This idea was realized by Zhou et al. (2011). In this work, a feedback set of a UE is defined as a subset of BSs to cooperatively serve the UE. Each UE optimizes its feedback set distributively according to the expected signal-to-noise-plus-interference ratio (SNIR) which includes the interference caused by the feedback error. The proposed algorithm outperforms the conventional nonadaptive CSI feedback.

The adaptive quantized CSI feedback deals with the limits on feedback capacity. The basic idea is to utilize the feedback bits on the “most valuable” channels. It is possible that no bits are allocated to some BS as the signal strength is too weak. In this case, the practical cluster size shrinks. Hence, considering the feedback capacity constraint, there is a trade-off between the cooperative gain with large cluster size and the interference caused by the feedback error. Again, the cluster size should not be too large in practice.

Feedback Scheduling with Constrained Backhaul

The backhaul link among BSs is responsible for sharing UEs’ data as well as CSI. When the backhaul link capacity is limited, not all the information can be shared. How to allocate the limited backhaul capacity should be carefully studied. The downlink achievable rate region under backhaul capacity constraint was characterized by Simeone et al. (2009). To fully exploit the advantage of BS cooperation, the idea of superposition coding can be introduced to enhance the ability of sharing data. One way is to share the integrated other-cell signal status, i.e., interference, via interference forwarding (Kim 2008) in the uplink or rate splitting (Maric et al. 2007) in the downlink. The other is to share part of UEs’ data among neighboring BSs for cooperative precoding or interference cancellation (Marsch and Fettweis 2008).

When jointly considering the CSI feedback constraint, the problem becomes more complicated. Both feedback and backhaul

constraints result in imperfect information for cooperation. If the feedback is perfect, the CSI will become imperfect due to limited backhaul. On the other hand, if the backhaul capacity is sufficient, limited feedback still causes interference by imperfect CSI. Therefore, it is necessary to jointly consider both effects. Marsch and Fettweis (2009) jointly considered limited feedback and backhaul constraints and evaluated some detection and coding mechanisms. Based on the analysis, how to adjust CSI feedback scheme under backhaul constraint is further investigated. Intuitively, according to different backhaul capacities, the cooperation mode can switch between joint transmission and interference suppression or partly joint transmission plus partly interference suppression. When the backhaul capacity is sufficient, BSs can fully cooperate to obtain cooperation gain. On the other hand, when the backhaul capacity is insufficient, only the interference information can be exchanged, and interference suppression is adopted.

In summary, the application of BS cooperation in real system encounters imperfect conditions, including limited computational capacity, synchronization issue, limited feedback capacity, and limited backhaul capacity. Candidate solutions have been given with three categories: clustering, feedback bits allocation, and mixed cooperation and interference management policy. These schemes enhance the feasibility of BS cooperation in practice. In the next section, key applications will be given.

Key Applications

The BS cooperation has been standardized in 3GPP (TR36.819 2012), termed as Coordinated Multipoint (CoMP). The imperfect conditions are considered in the standard. For instance, the cooperation clustering with cluster size 2 is suggested, the feedback policy includes explicit feedback and implicit feedback, and the CoMP categories include joint processing (fully cooperate), coordinated scheduling/beamforming

(partially cooperate), and so on. The techniques have been widely applied in 4G/LTE networks.

Recent progress and applications include BS cooperation under energy-efficiency constraint. For instance, in smart grid scenario, BSs need not only cooperate in signal domain but also in energy domain, so that the provided energy is sufficient for cooperation (Xu and Zhang 2015). In energy-harvesting scenario, where each BS is powered by renewable energy that is intermittent and imbalance, novel cooperation protocol is required to tackle the energy imbalance between cooperating BSs (Gong et al. 2016). The energy constraint can be viewed as another imperfect condition. Hence, the design under energy-efficiency constraint can be considered as an extension on the research of BS cooperation with imperfect conditions.

Cross-References

- [Coordinated Multipoint Transmission and Reception](#)
- [Key Technologies in 4G/LTE Network](#)
- [Network Slicing-Enabled Green C-RAN](#)

References

- Bhagavatula R, Heath RW (2011) Adaptive limited feedback for sum-rate maximizing beamforming in cooperative multicell systems. *IEEE Trans Signal Process* 59(2):800–811
- Gong J, Zhou S, Niu Z, Geng L, Zheng M (2011) Joint scheduling and dynamic clustering in downlink cellular networks. In: *IEEE global telecommunications conference (GlobeCom)*, pp 1–5
- Gong J, Zhou S, Zhou Z (2016) Networked MIMO with fractional joint transmission in energy harvesting systems. *IEEE Trans Commun* 64(8):3323–3336
- Huang H, Trivellato M, Hottinen A, Shafi M, Smith PJ, Valenzuela R (2009) Increasing downlink cellular throughput with limited network MIMO coordination. *IEEE Trans Wirel Commun* 8(6):2983–2989
- Karakayali MK, Foschini GJ, Valenzuela RA (2006) Network coordination for spectrally efficient communications in cellular systems. *IEEE Wirel Commun* 13(4):56–61
- Kim YH (2008) Capacity of a class of deterministic relay channels. *IEEE Trans Inf Theory* 54(3):1328–1329

- Liu J, Wang D (2009) An improved dynamic clustering algorithm for multi-user distributed antenna system. In: International conference wireless communication signal processing, pp 1–5
- Love DJ, Heath RW, Lau VKN, Gesbert D, Rao BD, Andrews M (2008) An overview of limited feedback in wireless communication systems. *IEEE J Sel Areas Commun* 26(8):1341–1365
- Maric I, Yates RD, Kramer G (2007) Capacity of interference channels with partial transmitter cooperation. *IEEE Trans Inf Theory* 53(10):3536–3548
- Marsch P, Fettweis G (2008) On base station cooperation schemes for downlink network MIMO under a constrained backhaul. In: IEEE global telecommunications conference (Globecom), pp 1–6
- Marsch P, Fettweis G (2009) On downlink network MIMO under a constrained backhaul and imperfect channel knowledge. In: IEEE global telecommunications conference (Globecom), pp 1–6
- Papadogiannis A, Bang HJ, Gesbert D, Hardouin E (2008a) Downlink overhead reduction for multi-cell cooperative processing enabled wireless networks. In: IEEE 19th international symposium personal, indoor and mobile radio communication (PIMRC), pp 1–5
- Papadogiannis A, Gesbert D, Hardouin E (2008b) A dynamic clustering approach in wireless networks with multi-cell cooperative processing. In: IEEE international conference communication (ICC), pp 4033–4037
- Saleh AAM, Rustako A, Roman R (1987) Distributed antennas for indoor radio communications. *IEEE Trans Commun* 35(12):1245–1251
- Simeone O, Somekh O, Poor HV, Shamai (Shitz) S (2009) Downlink multicell processing with limited-backhaul capacity. *EURASIP J Adv Signal Process* 2009(1):840–814
- TR36819 (2012) Coordinated multi-point operation for lte physical layer aspects (release 11). Technical report, 3GPP
- Xu J, Zhang R (2015) CoMP meets smart grid: a new communication and energy cooperation paradigm. *IEEE Trans Veh Technol* 64(6):2476–2488
- Zhang J, Andrews JG (2008) Distributed antenna systems with randomness. *IEEE Trans Wirel Commun* 7(9):3636–3646
- Zhang J, Andrews JG (2010) Adaptive spatial interference cancellation in multicell wireless networks. *IEEE J Sel Areas Commun* 28(9):1455–1468
- Zhang J, Chen R, Andrews JG, Ghosh A, Heath RW (2009) Networked MIMO with clustered linear precoding. *IEEE Trans Wirel Commun* 8(4):1910–1921
- Zhou S, Zhao M, Xu X, Wang J, Yao Y (2003) Distributed wireless communication system: a new architecture for future public wireless access. *IEEE Commun Mag* 41(3):108–113
- Zhou S, Gong J, Niu Z, Jia Y, Yang P (2009) A decentralized framework for dynamic downlink base station cooperation. In: IEEE global telecommunications conference (Globecom), pp 1–6
- Zhou S, Gong J, Niu Z (2011) Distributed adaptation of quantized feedback for downlink network mimo systems. *IEEE Trans Wirel Commun* 10(1):61–67

Basic Theory of Cyberspace Security

- [General Theory of Information Security](#)

Beam Division Multiple Access (BDMA) Transmission

- [Massive MIMO BDMA Transmission](#)

Beamforming

- [Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices](#)

Beamspace MIMO

- [Lens Antenna Array](#)

Behavioral Economics

- [Prospect Theory for Communication Networks](#)

Bid Rigging

- [Collusion](#)

Bidding

- [Auction](#)

Big Data Analyzing Systems

► Big Data Networks

Big Data Based Service Provisioning

► Data-Driven Service Provisioning

Big Data Computing

► Data-Driven Resource Allocation

Big Data in 5G

Yue Wang and Zhi Tian
ECE Department, George Mason University,
Fairfax, VA, USA

Synonyms

5G; Machine learning; Next-generation wireless networks; Sparse signal processing; Wireless Big Data

Definition

The fifth-generation wireless systems, abbreviated as 5G (Andrews et al. 2014), are proposed as the next wireless and mobile communications standards beyond the current 4G standards. 5G networks not only aim at providing higher data rate, lower latency, larger capacity, and better customer experience than 4G but also commit

to fulfilling the Internet of things (IoT) with reliable and secure services at low costs (Atzori et al. 2010). To this end, 5G networks call for and rely on seamless operations of distinctive wireless technologies and solutions, including cognitive radio (CR) (Akyildiz et al. 2006), massive multiple-input multiple-output (maMIMO) (Larsson et al. 2014), millimeter wave (mmWave) communications (Rappaport et al. 2013), heterogeneous network (HetNet) architecture, cloud-based radio access, edge computing and caching (Hu et al. 2015), device and interference management (Asadi et al. 2014), etc. These revolutionary technologies deal with very wide radio spectrum, extremely high-frequency bands, dynamic spectrum access, large-scale antenna arrays, a huge number of devices, massive connectivity, context-aware computing, and so on. All these issues signify signal and data processing in regimes of exceedingly large volume, size, and dimension. As 5G services and technologies both support and contribute to an unprecedented amount of data traffic, the field of wireless communications has entered a Big Data era with imminent challenges and opportunities.

In the era of data deluge, Big Data refer to the kinds of data that are extremely large in terms of the size of available dataset and/or extremely complicated in terms of the processing complexity to separate, classify, and analyze the data. Thus, management of Big Data goes beyond the capability of conventional signal processing techniques and data analysis tools. Further, Big Data in 5G present unique challenges due to the characteristics of wireless and mobile data traffic and service expectations. Multi-domain resources, e.g., time, frequency, space, energy, and code, are intertwined in complex ways, and real-time content delivery are often expected by lightweight mobile devices. As a result, traditional processing and analysis techniques become inadequate to deal with such a huge amount of complex data within a tolerable elapsed time at affordable costs. Fortunately, recent advances and developments in machine learning and sparse signal processing illuminate pathways to embrace the challenges and opportunities from Big Data in 5G.

Historical Background

With the pervasive connectivity offered by 5G technologies, it is becoming a reality that we live in an immersive smart world with smart cities, smart networks, smart cells, smart grids, smart homes, smart vehicles, smart industries, smart devices, smartphones, smart sensors, and so on. The rapidly growing 5G and IoT applications generate an unprecedentedly vast amount of diverse data, which create and expand the regime of Big Data in 5G. On the one hand, the huge volume and high complexity of Big Data cannot be adequately and efficiently supported by legacy wireless technologies in spectrum access, resource allocation, and network management, all of which need to be revamped to address the Big Data challenges. On the other hand, Big Data, as new fuel for learning and predicting channel and network traffic behavior, offer great opportunities for improved system performance and much enhanced quality of experience in 5G.

In the regime of Big Data in 5G, although the dimension of signals and the quantity of data grow unprecedentedly large, there exist unique opportunities because of some useful properties, structures, and features exhibited in the underlying signals and data at both the signal and network levels. Harnessing such properties, structures and features not only allow to acquire data efficiently and process signals smartly but also enable prediction, control, management, and optimization of 5G networks in a proactive manner. Along this line, recent advances in powerful machine learning and intelligent signal processing techniques play key roles in not only supporting Big Data applications but also developing revolutionary 5G technologies in the Big Data era.

At the network level, powerful machine learning techniques have been employed to learn data traffic patterns in 5G networks for efficient network management and control. Machine learning, as a promising technology in artificial intelligence, gives machines the ability to learn without following strictly static program instructions by making predictions or decisions (Alpaydin 2004).

Generally speaking, machine learning enables a machine to learn the execution of some class of tasks T from experience E with respect to certain performance measure P , where the performance achieved at tasks in T is measured by P and can be improved with experience E . Machine learning approaches are typically classified into three broad categories, depending on the nature of the dataset for learning or the feedback available to a learning system. They are supervised, unsupervised, and reinforcement learning, where supervised and unsupervised learning indicate whether or not there are labeled samples in the input dataset and reinforcement provides feedback in terms of rewards and punishment to navigate the learning process. Recently, both the increase of available data volume and the improvement on hardware computational ability have contributed to the fast development of a new concept of deep learning (LeCun et al. 2015), which makes use of multiple hidden layers in an artificial neural network. This technique aims to mimic the way that the human brain processes light and sound into vision and hearing (LeCun et al. 2015).

At the signal level, intelligent signal processing techniques have been developed to exploit useful structural features of the complex channels and spectra in 5G systems for effective data acquisition and transmissions. Facing the large size and high dimension of sensing problems arising in wireless communications, compressive sensing (CS) (Candes and Wakin 2008), a.k.a. compressed sensing or compressive sampling, provides a new paradigm of simultaneous data acquisition and compression to effectively reduce the sampling costs of high-dimensional signals by utilizing the fact that typical signals of interest are often sparse in a certain domain via projection onto certain known basis. It is the sparsity of signals that enables CS techniques to reconstruct signals from far fewer samples than those required by the Nyquist sampling theory. As Big Data applications continue to grow in size and number, it is critical to deal only with measurements of data that are informative for specific inference tasks of interest, in order to limit the required sensing cost, as well as the related costs of storing or communicating Big Data. Alterna-

tive to CS for deterministic signals, a novel statistical signal processing framework, called compressive statistics sensing (CSS), is developed to leverage the statistical structure of random processes to enable compression (Romero et al. 2016). Capitalizing on parsimonious representations, CSS technology allows compression and reconstruction tasks to be addressed in broader applications, such as wideband spectrum sensing, frequency detection, direction-of-arrival estimation, power spectrum estimation, incoherent imaging, etc.

Foundations and Applications

When entering the era of wireless Big Data, it is of crucial importance to demonstrate how advances in artificial intelligence and signal processing can be used to overcome challenges and seek opportunities in future 5G wireless networks. Specific exemplary scenarios below describe some situations and applications in which different kinds of machine learning and sparse signal processing techniques play key roles in utilizing Big Data for wireless networks.

Supervised Learning

Supervised learning refers to the machine learning task of inferring an input-output mapping function from labeled training data, where the dataset contains observations of both the input objects and the resulting output values. Specific techniques include regression analysis, k -nearest neighbors (KNN) algorithm, support vector machine (SVM), and Bayesian learning (Alpaydin 2004). The regression analysis relies on a set of statistical processes for estimating the relationships among variables, which aims at predicting the value of one or more estimation targets. The estimation target is a function of the independent variables. The KNN and SVM algorithms are mainly used for classification. KNN classifies an object into a category based on a majority vote of the object's neighbors. That is, the object is assigned into a class that is most common among its k -nearest neighbors. In contrast, the SVM algorithm resorts to nonlinear

mapping, which transforms the original training data into a higher dimension where the data become separable. Then, SVM searches for the optimal linear hyperplane that can separate one class from another in the higher dimension. Bayesian learning is to compute and maximize the a posteriori probability distribution of a target variable conditioned on both the input signals and all of the training instances. Some generative models that can be learned with the aid of Bayesian techniques include the Gaussian mixture model, expectation maximization, and hidden Markov model.

In wireless communications, supervised learning can be used for training-based estimation and prediction of radio parameters. For example, in channel estimation for nonlinear MIMO systems where nonlinearities are present in either the transmitter or the receiver sides, a multivariate regression framework is modeled based on the SVM technique to utilize the MIMO channel multidimensionality (Sanchez-Fernandez et al. 2004). The use of the multidimensional regression enables to exploit the dependencies in the channel and then make each estimate less vulnerable to the added noise. In solving such a nonlinear MIMO channel estimation problem, an iterative reweighted least-squares algorithm is developed for the regression of multiple variables. Thanks to the benefits of SVM in solving nonlinear problems, this SVM-based channel estimation method can take advantage of the MIMO spatial diversity and is able to discover the dependencies between the transmitted and received signals.

For the link adaptation problem in MIMO-OFDM WiFi systems, supervised machine learning has also been successfully applied. In Daniels et al. (2010), the feature space is first defined to characterize the link quality by ordering the post-processing signal-to-noise ratio (SNR) and selecting those with high values as the most relevant link-level performance indicators. Then, the second step is to identify a mapping function between the defined feature set and the link configuration to fulfill the design objective. The mapping task is treated as a classification problem to determine the relationship between the channel

state and the configuration for adaptive modulation and coding, where KNN can be adopted as a classifier.

Another fruitful application is context-aware energy enhancement for mobile devices in IoT applications (Donohoo et al. 2014). KNN techniques have been applied to learn and utilize the spatiotemporal and device context to predict the device wireless data and location interface configurations that can optimize energy consumption in mobile devices (Donohoo et al. 2014). The experiment results in Donohoo et al. (2014) show that the successful rate of energy demand prediction exceeds 90% with the use of KNN algorithms.

Furthermore, Bayesian learning techniques can be used for learning and predicting spectral characteristics in cognitive radio systems and channel estimation in massive MIMO. For example, to solve the pilot contamination problem in massive MIMO systems, Wen et al. (2015) suggest to estimate both the channel parameters of the desired links in a target cell and those of the interfering links in other adjacent cells, in which channel estimation is carried out with the aid of Bayesian learning techniques. For cognitive radio networks, a cooperative wideband spectrum sensing approach is proposed to detect primary users via the expectation maximization algorithm (Assra et al. 2016). In Choi and Hossain (2013), a two-state hidden Markov process is designed to learn the spectrum occupancy of the primary user, whose presence and absence are mapped to a two-state observation space for Bayesian learning.

Unsupervised Learning

Unsupervised learning is to fulfill the learning task of inferring a function to describe the underlying structure or distribution from unlabeled data, which means the desired output values (e.g., classification or categorization outcomes) are not included in the observations (Alpaydin 2004). Since the data given to the learner are unlabeled, there is no evaluation of the accuracy of the learned structure as the output of the learning algorithm. As an example of unsupervised learning, k -means clustering is used for cluster analy-

sis in data mining, which aims to partition observations into k clusters where each observation belongs to the cluster with the nearest mean. The k -means clustering algorithm runs in an iterative manner. In each iteration, an object is assigned to the cluster whose centroid is the nearest one to the object and then each cluster and its centroid are updated by minimizing the in-cluster differences. The iterations will terminate until convergence is reached.

For wireless heterogeneous networks (HetNet) with densely deployed small cells, clustering is a common yet challenging task given the diverse types of operating networks/protocols and varied cell sizes. For example, small cells have to be carefully clustered to mitigate intercell interference, and terminal users also have to be clustered dynamically to avoid frequent cell handovers and achieve optimal offloading. In fact, the handover performance and intercell interference coordination critically determine the network efficiency and user experience. For cell and user clustering, several unsupervised learning methods can be used, such as k -mean clustering and principal component analysis (PCA). PCA can be viewed as a relaxed solution to k -means clustering, where the PCA subspace spanned by the principal directions is identical to the cluster centroid subspace. PCA transforms a set of potentially correlated variables into a set of uncorrelated variables a.k.a. the principal components.

In contrast to PCA, independent component analysis (ICA) is a statistical technique for revealing hidden factors that underlie sets of data variables. Based on these underlying factors, ICA decomposes the data variables into a set of independently additive components. This is done by assuming that the subcomponents are non-Gaussian and mutually independent from each other. ICA is useful for blind source separation. In Qiu et al. (2011), ICA is used in a CR-based smart grid system to recover the simultaneous wireless transmissions of smart utility meters, where the power utility station applies ICA to blindly separate the signals received from all the smart meters before the signals can be decoded, by utilizing the statistical properties of the signals.

Reinforcement Learning

Reinforcement learning is inspired by behaviorist psychology to deal with learning tasks when the feedback of the instantaneous reward is available. It is concerned with how learning agents ought to take actions in an environment based on feedback, in order to maximize the cumulative reward over time (Otterlo and Wiering 2012). The environment in reinforcement learning is typically formulated as a Markov decision process (MDP). The main differences between the classical inference techniques and reinforcement learning algorithms include: (1) the reinforcement learning techniques do not request any prior knowledge about the MDP, and (2) they target large-scale MDPs where classical methods become infeasible. Reinforcement learning focuses on the online performance, which enables to reach a balance between exploration of uncharted territory and exploitation of obtained knowledge. Such a trade-off can be achieved via the multiarmed bandit (MAB) and finite MDPs (Kaelbling et al. 1996). MDP describes a discrete-time stochastic control process and provides a framework for modeling the decision-making situations in which the outcomes are partly random and partly under the control of a decision-maker. At each step, the MDP stays in a certain state, and the decision-maker may choose any legitimate action that is available in the current state. At the following step, the MDP randomly moves into a new state and gives the decision maker a corresponding reward. The probability that the MDP moves into its new state is influenced by both the selected action and the system's inherent transitions. In this sense, the new state depends on the current state and the action selected by the decision-maker. Meanwhile, the state transition probability is conditionally independent of all previous states and actions, which makes the MDP satisfy the Markov property. MDP usually assumes that the state is known when action is to be taken. However, if this assumption is not guaranteed, then the MDP problem becomes partially observable, a.k.a. POMDP. The POMDP models an agent decision process where the system dynamics are determined by an MDP, although the agent is unable to directly observe the underlying state

and only has partial knowledge. Thus, the agent must maintain a probability distribution of the possible states based on a set of observations and observation probabilities and the underlying MDP. In the learning task for the MAB problem, an agent seeks to simultaneously perform exploration and exploitation: the former is to acquire new knowledge, and the latter is to optimize the decisions based on existing knowledge.

For decision-making in 5G wireless systems, the terminal users in a network can be regarded as agents, and the network constitutes the environment of the MDP/POMDP models. The applications of reinforcement learning include transmit power control in energy-constrained systems, channel selection in device-to-device (D2D) communications, dynamic spectrum access in CR networks, and so on. For example, future IoT systems may feature in wireless energy harvesting sensors that operate using energy harvested from environmental sources such as the sun, wind, vibrations, etc. For these sensors, energy management is crucial to ensure continuous and reliable operation with minimized power outage. In Aprem et al. (2013), the problem of transmit power control with packet retransmissions is formulated as making a sequential decision to achieve optimal outage probability performance. Through acknowledgement (ACK) and negative-acknowledgement (NACK) feedback messages, the channel information is implicitly provided to the sensors, which can be exploited in deciding the transmit power level for subsequent transmission attempts. Since the ACK and NACK messages only reveal partial channel information to the sensors, this problem is cast as a POMDP formulation. Accordingly, computationally efficient solutions have been developed based on the maximum likelihood criterion and the voting heuristic policy (Aprem et al. 2013).

In the context of D2D communications, distributed and autonomous channel selection is considered as an underlay to a cellular network. D2D users directly communicate with each other by exploiting the cellular spectrum, while

their individual decisions are not governed by any centralized controller. Self-interested D2D users competing for access to optimize their own performance form a distributed system, where the transmission performance actually relies on channel availability and quality. However, such information is hard to acquire for autonomous D2D users. Further, the adverse impact of D2D communications to cellular systems should be minimized. Under these limitations, a network-assisted distributed channel selection approach is proposed by modeling a multiplayer (MP) MAB game with broadcast side information indicating channels selected by other D2D users (Maghsudi and Stanczak 2015). In this MP-MAB game, each D2D user is viewed as a player, and the channels are regarded as multiple arms, where selecting a channel corresponds to pulling an arm. Then, a distributed solution is developed by combining no-regret learning and calibrated forecasting for multiplayer stochastic learning problems. Similarly, in CR scenarios, the problem of dynamic spectrum access of decentralized secondary CR users can also be modeled as a MP-MAB game (Liu and Zhao 2010), where each CR user independently searches for spectrum opportunities without exchanging sensing information with others.

Deep Learning

Deep learning refers to the learning techniques that allow computational models to take on multiple processing layers in order to learn representations of data with multiple levels of abstraction (LeCun et al. 2015). Deep learning has led a new trend in artificial intelligence by providing a powerful set of techniques for learning via deep neural networks (DNN). In DNN, connections and interactions of neurons from layer to layer contribute to feature extraction and transformation, where each layer uses the output from the previous layer as the current input. Such a topology allows a biologically inspired programming paradigm where an agent is able to learn from observational data via representing and utilizing the abstract features among data (Schmidhuber 2015). By continuously learning

multilevel features of the data, higher-level features are learned and formulated from lower-level features to form a hierarchical representation. In this way, deep learning translates the features of data into compact intermediate representations over layered structures, which can be used to remove redundancy and detect saliency. In the era of 5G, wireless traffic and user data are experiencing a tremendous growth due to pervasive devices, ubiquitous connections, and diverse services and applications. Different from the centralized and static architecture of conventional cellular networks, future wireless networks and systems are moving toward decentralized and flexible network architectures. In this context, deep learning provides useful tools to acquire underlying trends, predict upcoming events, and harvest correlations and statistical probabilities, given a huge amount of available data for training. Such efforts enable 5G wireless networks to make proactive decisions such as proactive resource allocation by taking advantage of context awareness and edge computing and caching (Wang et al. 2014), which thereby improve network performance and efficiency. In return, these 5G techniques offer pervasive network connectivity for data collection and transmission, which make it possible to enhance Big Data analytic tools and deep learning techniques within 5G networks for accurate content popularity estimation and prediction. After traffic content learning, popular contents are cached in the intermediate servers so that demands from users for the same content can be accommodated easily without duplicate transmissions from remote servers (Bastug et al. 2014). As a result, redundant traffic can be significantly eliminated to save precious wireless network resources.

In addition, given the available large volume of user and network data, Big Data analytics are inherently synergistic with the trend of software-defined networking (SDN) in 5G (Haleplidis et al. 2015), which refers to network programmability by initializing, controlling, managing, and updating network behavior dynamically via open interfaces. While SDN has emerged as a promising networking solution, opening up interfaces gives rise to the risk of security

threats to both user and network data. To solve such problems, neural network can be adopted for anomaly detection in SDN environments. In Jadidi et al. (2013), to protect a network from attacks, a flow-based anomaly detection algorithm is developed based on a multilayer perceptron and gravitational search scheme, in which a neural network uses a large volume of traffic data to classify benign and malicious flows with high accuracy.

Wideband Spectrum Sensing

At the signal level, Big Data challenges arise primarily from the large size and high dimension of transmitted and received signals that 5G systems have to cope with. For instance, cognitive radio (CR) has been acknowledged as an important technology for dynamic spectrum access in future 5G, which aims to overcome the dilemma between radio frequency (RF) scarcity and spectrum underutilization. The first and key task in CR is to perform accurate and fast spectrum sensing in order to determine the spectrum occupancy of existing (primary) users and identify potential transmission opportunities for (secondary) CR users. To harvest the benefits of open spectrum access and dynamic spectrum sharing, spectrum sensing is usually conducted over a very wide range of frequency bands. In the wideband regime, a major challenge stems from the high costs in RF signal acquisition, which hinders the implementation of fast and accurate spectrum detection. On the other hand, the spectrum underutilization over most assigned frequency bands induces widespread sparsity in the frequency domain. Capitalizing on such sparseness of the underutilized spectrum in open-access networks, CS techniques have been tailored and applied for wideband spectrum sensing using a much smaller number of samples than that required by the Nyquist sampling rate (Tian and Giannakis 2007; Polo et al. 2009). This is because the sampling rate in CS is dictated by the sparsity order of the signals, rather than the large dimension of the original signals. Since the average spectrum utilization of most frequency bands is below 20% at a given time and location (Akyildiz et al. 2006), the CS approach allows for sampling

at a small fraction of the Nyquist rate without loss of sensing accuracy.

Considerable research efforts have been devoted to realizing the potential of CS-based sensing and sparse signal processing in wireless CR systems under practical operating scenarios. For instance, in practical wireless environments, the actual sparsity order of wideband spectrum is usually unknown a priori or even dynamically changing over time. To circumvent this problem, a concept of sparsity order estimation is proposed in Wang et al. (2012a), which can be applied to design a two-step compressive spectrum sensing solution for wideband CR (Wang et al. 2010). Wireless fading constitutes a major performance-degrading factor to sensing techniques in practice. To cope with the fading effects, collaborative spectrum sensing among distributed CR users provides an effective way to improve the sensing performance (Zeng et al. 2010). For wideband collaborative CR systems, to jointly collect both performance gain and complexity gain, a cooperative spectrum sensing solution is designed based on matrix rank minimization techniques (Wang et al. 2011). Therein the spectrum measurements of all collaborative CR users are modeled to possess a decent low-rank property. Accordingly, a nuclear norm minimization problem is formulated to jointly identify the nonzero support and hence the overall wideband spectrum occupancy, in which the low-rank property enables efficient utilization and desired tradeoff between detection diversity and sampling costs (Wang et al. 2012b).

Under the conventional centralized network architecture for spectrum sensing, although the sensing performance can be globally optimal, the energy costs in reporting local information to the fusion center and conveying global decisions back to the CR users can be undesirably high. Alternatively, decentralized cooperative spectrum sensing becomes attractive for its low communication overhead and robustness to node and link failure. Decentralized schemes have been developed based on consensus techniques (Bazerque and Giannakis 2010; Tian 2008). These algorithms adopt a consensus averaging technique, where the average value of all the local spectrum

decisions is computed in a decentralized manner and taken as the global decision.

The framework of consensus-based collaborative sensing has been extended to address the technical hindrance in practical multi-hop CR networks. In a CR network adopting dynamic access, there is no guarantee on stringent synchronization to ensure that all CR users stay silent during the sensing stage. Further, due to the multi-hop nature, a CR user during the sensing stage may be subject to distinct spectral emissions from sources within its local area, such as other coexisting CR users in the transmission mode, mistimed cooperative CR users in the sensing mode, or local interference. These individually received spectral components, termed as spectral innovations, are sparse but complicate the cooperative task of detecting the common spectrum support of primary users. To deal with this complex but practical problem, decentralized multi-hop cooperative spectrum sensing solutions have been developed in Zeng et al. (2010) and Zeng et al. (2011), in which each CR user estimates the common spectrum of primary users and its own local spectral innovation in an alternating manner and exchanges proper information with neighboring CRs to reach global fusion and consensus on the estimated primary users' spectrum occupancy.

Besides the wide bandwidth and wireless fading, noise uncertainty raises another critical issue that degrades the performance of spectrum sensing. Cyclic feature detection as a kind of statistical signal processing technique works robust under noise uncertainty and low SNR (Gardner 1991) but requires high-rate sampling which is very costly especially in the wideband regime. To solve this problem, a compressive cyclic feature detection technique has been developed for wideband spectrum sensing by exploiting the unique sparsity property of the two-dimensional cyclic spectra of cyclostationary signals (Tian 2011). Along this line, a new compressed covariance sensing framework is proposed for extracting useful second-order statistics of wideband random signals from digital samples taken at sub-Nyquist rates (Romero et al. 2016; Tian et al. 2012).

Sparse Channel Estimation

To meet the growing demands for high-rate and large-capacity wireless services, massive multiple-input multiple-output (MIMO) technology and millimeter wave (mmWave) communications have emerged as a natural pair to provide a promising option for 5G networks. In MIMO communications, the channel state information (CSI) is required via channel estimation for transmitter design and receiver demodulation. However, the large number of antennas in massive MIMO systems results in a large-size channel matrix, which is costly to estimate, because traditional channel estimation techniques would consume heavy resources in terms of training symbols and CSI feedback. To reduce such heavy training overhead, CS has been advocated for CSI estimation in mmWave MIMO systems (Bajwa et al. 2010; Schniter and Sayeed 2014; Alkhateeb et al. 2014; Gao et al. 2016; Wang et al. 2016a, b). By exploiting the sparse structure of mmWave MIMO channels due to the limited multipath scattering, the CS-based approach can estimate the CSI from a relatively small set of compressively collected training symbols. In Bajwa et al. (2010) and Schniter and Sayeed (2014), the task of estimating sparse multipath channels is formulated as a sparse vector recovery problem by vectorizing the sparse channel matrix and representing it within a three-dimensional angle-delay-Doppler space. Subsequently, sparse signal recovery techniques can be applied to acquire the vectorized sparse channel from compressive measurements within a short sensing time (Schniter and Sayeed 2014; Gao et al. 2016). In Alkhateeb et al. (2014), an adaptive CS-based algorithm is proposed to estimate the sparse channel with hybrid analog/digital hardware architecture. In Wang et al. (2016a), a diagonal-search greedy pursuit algorithm is developed to estimate the second-order statistics of the sparse MIMO channel. To reduce the sensing time and computational complexity caused by the vectorization operation, a fast channel estimation solution is developed in Wang et al. (2016b), which directly decouples the original channel estimation problem into three reduced-size subproblems, i.e., angle-of-arrival

(AoA) estimation, angle-of-departure (AoD) estimation, and fading coefficient estimation, which can be solved in a sequential manner.

The CS-based channel estimation techniques for mmWave MIMO systems implicitly rely on the assumption that the values of the AoD/AoA of sparse paths fall on some known grid. However in practice, the AoD/AoA can be continuously valued off the grid, in which case the CS-based methods may suffer from the power leakage effect due to basis mismatch of the on-grid assumption (Chi et al. 2011). To circumvent the on-grid assumption, gridless channel estimation solutions can utilize the two-level Toeplitz structure of mmWave massive MIMO channel by applying two-dimensional (2D) atomic norm minimization (ANM) to estimate the continuous-valued channel state information with super resolution. The 2D ANM is a two-level structure-based optimization approach where the Vandermonde structure of the transmit/receive array manifold is captured by a two-level Toeplitz matrix, which can be constructed from the received array data via semidefinite programming (SDP). Such a gridless channel estimation approach attains super-resolution AoD/AoA estimation accuracy within a very short sensing time, possibly from measurements collected at a single-time snapshot.

Noticeably, the computational complexity of the SDP-based ANM formulation is decided by the numbers of transmit and receive antennas used for MIMO channel estimation. As a result, the computational complexity goes up quickly as the antenna size increases and may become intractable for terminal devices. To reduce the high complexity, an efficient truncated 2D gridless channel estimation solution is proposed for mmWave massive MIMO systems (Wang et al. 2017). It reformulates the original full-size ANM to a truncated ANM (T-ANM) version by activating a small part of transmit/receive antennas. The T-ANM approach is motivated from an important fact that the mmWave MIMO channel provides not only the two-level Toeplitz structure but also a salient low-rank property due to the limited scattering propagation at mmWave frequency. Further complexity reduction is accomplished by

introducing an alternative form of ANM, termed decoupled ANM (D-ANM) (Tian et al. 2017). It is another optimal ANM formulation for gridless 2D harmonic retrieval problem, which adopts a new matrix form set of atoms to naturally decouple the joint observations in both AoA and AoD dimensions without loss of optimality. The D-ANM strategy reformulates the original large-scale 2D problem into two one-level Toeplitz matrices, which can be solved by simple 1D estimation with automatic pairing. Therefore, D-ANM dramatically reduces the constraint size in SDP-based ANM formulation, thus greatly reducing the overall computational complexity.

Summary and Outlook

This article reviews emerging technologies and advances achieved in both machine learning and sparse signal processing and focuses on data analytic tools that can be used in future 5G wireless networks. To demonstrate their broad applications in various wireless systems, selected examples are illustrated, including CR, HetNet, mmWave, massive MIMO, software-defined networking, etc.

There are still a range of open questions in Big Data analytics for 5G wireless networks. For instance, it is quite challenging to apply machine learning to gain full awareness of the RF environments over space, time, frequency, and location, due to the dynamic nature and the lack of labeled data. As opposed to computer vision and speech recognition where the output of machine cognition can be readily compared and verified against human visual and auditory perception, no such option is available for radio signals.

Cross-References

- [5G Wireless](#)
- [Cellular Unlicensed Spectrum Technology](#)
- [Cognitive Heterogeneous Networks](#)
- [Machine Learning in Wireless Sensor Networks for the Internet of Things](#)

- Mobile Edge Caching in HetNets
- Network Selection in Heterogeneous Wireless Networks
- Wireless Edge Caching
- Wireless Fading Channel Model for 5G and Future

References

- Akyildiz I, Lee W, Vuran M, Mohanty S (2006) NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey. *Comput Netw* 50(13):2127–2159
- Alkhateeb A, El Ayach O, Leus G, Heath RW (2014) Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE J Sel Topics Signal Process* 8(5):831–846
- Alpaydin E (2004) Introduction to machine learning. MIT Press, Cambridge
- Andrews G et al (2014) What will 5G be? *IEEE J Sel Areas Commun* 32(6):1065–1082
- Aprem A, Murthy CR, Mehta NB (2013) Transmit power control policies for energy harvesting sensors with retransmissions. *IEEE J Sel Topics Signal Process* 7(5):895–906
- Asadi A, Wang Q, Mancuso V (2014) A survey on device-to-device communication in cellular networks. *IEEE Commun Surv Tutor* 16(4):1801–1819
- Assra A, Yang J, Champagne B (2016) An EM approach for cooperative spectrum sensing in multiantenna CR networks. *IEEE Trans Veh Technol* 65(3):1229–1243
- Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. *Comput Netw* 54(15):2787–2805
- Bajwa WU, Haupt J, Sayeed AM, Nowak R (2010) Compressed channel sensing: a new approach to estimating sparse multipath channels. *Proc IEEE* 98(6):1058–1076
- Bastug E, Bennis M, Debbah M (2014) Living on the edge: the role of proactive caching in 5G wireless networks. *IEEE Commun Mag* 52(8):82–89
- Bazerque JA, Giannakis GB (2010) Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Trans Signal Process* 58(3):1847–1862
- Candes EJ, Wakin MB (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25(2):21–30
- Chi Y, Scharf LL, Pezeshki A, Calderbank R (2011) Sensitivity to basis mismatch in compressed sensing. *IEEE Trans Signal Process* 59(5):2182–2195
- Choi KW, Hossain E (2013) Estimation of primary user parameters in cognitive radio systems via hidden Markov model. *IEEE Trans Signal Process* 61(3):782–795
- Daniels RC, Caramanis CM, Heath RW (2010) Adaptation in convolutionally coded MIMO-OFDM wireless systems through supervised learning and SNR ordering. *IEEE Trans Veh Technol* 59(1):114–126
- Donohoo BK et al (2014) Context-aware energy enhancements for smart mobile devices. *IEEE Trans Mob Comput* 13(8):1720–1732
- Gao Z, Hu C, Dai L, Wang Z (2016) Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels. *IEEE Commun Lett* 20(6):1259–1262
- Gardner W (1991) Exploitation of spectral redundancy in cyclostationary signals. *IEEE Signal Process Mag* 8(2):14–36
- Haleplidis E et al (2015) Software-defined networking (SDN): layers and architecture terminology. IRTF
- Hu Y et al (2015) Mobile edge computing: a key technology towards 5G, ETSI white paper
- Jadidi Z, Muthukkumarasamy V, Sithirasenan E, Sheikhan M (2013) Flow-based anomaly detection using neural network optimized with gsa algorithm. In: *IEEE 33rd international conference on distributed computing systems workshops*, Philadelphia, 8–11
- Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. *J Artif Intell Res* 4:237–285
- Larsson EG, Edfors O, Tufvesson F, Marzetta TL (2014) Massive MIMO for next generation wireless systems. *IEEE Commun Mag* 52(2):186–195
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444
- Liu K, Zhao Q (2010) Distributed learning in cognitive radio networks: multi-armed bandit with distributed multiple players. In: *IEEE ICASSP conference*, Dallas, 14–19 Mar 2010
- Maghsudi S, Stanczak S (2015) Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. *IEEE Trans Wirel Commun* 14(3):1309–1322
- Otterlo M, Wiering M (2012) Reinforcement learning and Markov decision processes. In: *Reinforcement learning*. Springer, Berlin/Heidelberg, pp 3–42
- Polo Y, Wang Y, Pandharipande A, Leus G (2009) Compressive wide-band spectrum sensing. In: *IEEE ICASSP conference*, Taipei, 19–24 Apr 2009
- Qiu RC et al (2011) Cognitive radio network for the smart grid: experimental system architecture, control algorithms, security, and microgrid testbed. *IEEE Trans Smart Grid* 2(4):724–740
- Rappaport TS et al (2013) Millimeter wave mobile communications for 5G cellular: it will work. *IEEE Access* 1(1):335–349
- Romero D, Ariananda D, Tian Z, Leus G (2016) Compressive covariance sensing: structure-based compressive sensing beyond sparsity. *IEEE Signal Process Mag* 33(1):78–93
- Sanchez-Fernandez M, de-Prado-Cumplido M, Arenas-Garcia J, Perez-Cruz F (2004) SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Trans Signal Process* 52(8):2298–2307
- Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117
- Schniter P, Sayeed AM (2014) Channel estimation and precoder design for millimeter-wave communications:

- the sparse way. In: Asilomar conference on signals, systems, and computers, Pacific Grove, 2–5 Nov 2014
- Tian Z (2008) Compressed wideband sensing in cooperative cognitive radio networks. In: IEEE GLOBECOM conference, New Orleans, 30 Nov–4 Dec 2008
- Tian Z (2011) Cyclic feature based wideband spectrum sensing using compressive sampling. In: IEEE ICC conference, Kyoto, 5–9 June 2011
- Tian Z, Giannakis GB (2007) Compressed sensing for wideband cognitive radios. In: IEEE ICASSP conference, Honolulu, 15–20 Apr 2007
- Tian Z, Tafesse Y, Sadler BM (2012) Cyclic feature detection from sub-Nyquist samples for wideband spectrum sensing. *IEEE J Sel Topics Signal Process* 6(1):58–69
- Tian Z, Zhang Z, Wang Y (2017) Low-complexity optimization for two dimensional direction-of-arrival estimation via decoupled atomic norm minimization. In: IEEE ICASSP conference, New Orleans, 5–9 Mar 2017
- Wang Y, Tian Z, Feng C (2010) A two-step compressed spectrum sensing scheme for wideband cognitive radios. In: IEEE GLOBECOM conference, Miami, 6–10 Dec 2010
- Wang Y, Tian Z, Feng C (2011) Cooperative spectrum sensing based on matrix rank minimization. In: IEEE ICASSP conference, Prague, 22–27 May 2011
- Wang Y, Tian Z, Feng C (2012a) Sparsity order estimation and its application in compressed spectrum sensing for cognitive radios. *IEEE Trans Wirel Commun* 11(6):2116–2125
- Wang Y, Tian Z, Feng C (2012b) Collecting detection diversity and complexity gain in cooperative spectrum sensing. *IEEE Trans Wirel Commun* 11(8):2876–2883
- Wang X et al (2014) Cache in the air: exploiting content caching and delivery techniques for 5G systems. *IEEE Commun Mag* 52(2):131–139
- Wang Y, Tian Z, Feng S, Zhang P (2016a) Efficient channel statistics estimation for millimeter-wave MIMO systems. In: IEEE ICASSP conference, Shanghai, 20–25 Mar 2016
- Wang Y, Tian Z, Feng S, Zhang P (2016b) A fast channel estimation approach for millimeter-wave massive MIMO systems. In: IEEE GlobalSIP conference, Washington, 7–9 Dec 2016
- Wang Y, Xu P, Tian Z (2017) Efficient channel estimation for massive MIMO systems via truncated two-dimensional atomic norm minimization. *IEEE ICC Conf*, Paris, 21–25 May 2017
- Wen C et al (2015) Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning. *IEEE Trans Wirel Commun* 14(3):1356–1368
- Zeng YH, Liang YC, Hoang AT, Zhang R (2010) A review on spectrum sensing for cognitive radio: challenges and solutions. *EURASIP J Adv Signal Process* 2010:1–15
- Zeng F, Tian Z, Li C (2010) Distributed compressive wideband spectrum sensing in cooperative multi-hop cognitive networks. *IEEE ICC Conf.*, 23–27 May 2010, Cape Town, South Africa
- Zeng F, Li C, Tian Z (2011) Distributed compressive spectrum sensing in cooperative multi-hop wideband cognitive networks. *IEEE J Sel Topics Signal Process* 5(1):37–48

Big Data in Networks

► Big Network Data

Big Data Networking Infrastructure

► Big Data Networks

Big Data Networks

Shigeng Zhang

School of Computer Science and Engineering,
Central South University, Changsha, P. R. China

Synonyms

[Big data analyzing systems](#); [Big data networking infrastructure](#)

Definition

Big data networks provide a distributed infrastructure for storing, transmitting, and processing a huge volume of data to extract useful information from the data.

Historical Background

The world is generating data much more fast than before. It is predicted that the total data generated in the world doubles every 2 years (Gantz and Reinsel 2012). The huge volume of data appears in various domains and applications, e.g., large graph or complex graph for social networks (Nie et al. 2017; Strogatz 2001), protein-protein interaction (PPI) networks (Stelzl et al. 2005) in bioinformatics, huge number of searching requests in searching engine, and the big data generated from Internet of Things (IoT) devices (Sun et al. 2016).

How to efficiently store and process the huge volume of data is an important technique issue. The idea of leveraging the resources from a set of computers to enhance processing capability and to manipulate a large volume of data can be traced back to early distributed computing systems in 1970s and 1980s (Gligor and Shattuck 1980). High-performance computing (HPC) (Dowd et al. 1998) and grid computing (Chervenak et al. 2000) played important roles in processing large volume of data since 1990s, and they are still important platforms to perform computational-aware tasks nowadays.

In 2006, Amazon released its elastic computer cloud (EC2) platform (Buyya et al. 2008), which provides a platform to aggregate resources from a set of powerless computers to obtain the required capability to process a large volume of data. This is different from previous approaches based on HPC. In the same year, Google released MapReduce (Dean and Ghemawat 2008) and Bigtable (Chang et al. 2008), which support parallel processing of big data in a cluster containing a large number of computers and managing a huge number of structured data in a cluster. The Spark framework (Zaharia et al. 2010) improves efficiency by using in-memory computation. Moreover, Spark supports iterative processing which are necessary in many machine learning tasks.

Foundations

Big Data VS. Networks

When the volume of data becomes huge, the data cannot be stored and processed on a single computer, and we have built networking systems to store, transmit, and process the big data. First, as the data exceeds the capability of a single computer node, it is necessary to build a network to connect a large number of nodes storing the data and exchange the data between them efficiently. This stimulates the development of data centers. Second, to process the big data, we need new network computing paradigms to process the data quickly and in parallel, such as MapReduce and Spark.

Big Data Storage

The huge volume of data are usually stored in *data centers* (DCs) (Al-Fares et al. 2010; Bari et al. 2013). A data center (Bari et al. 2013) is an infrastructure that consists of servers, storage and network devices, and auxiliary facilities such as power distribution systems and cooling systems. The communication infrastructure used in a DC is called a data center network (DCN). In order to improve the efficiency for big data storage and support fast processing of big data, current research in data center networks mainly focus on the following aspects:

- *Topology design:* The topologies used in DCNs can be classified into two categories: switch-centric topology and server-centric topology. In switch-centric DCN topologies, the network connection and routing functions are performed on switches, while the servers are mainly used for data storage and computation. Typical switch-centric topologies include Fat-Tree and VL2. In contrast, in server-centric topologies, the function of interconnection and routing are performed on servers, while the switches provide only simple crossbar functions. Typical server-centric topologies include Dcell, BCube, and uFdx.
- *TCP optimization:* Due to the special features of DCNs that are different from traditional Internet, performance of traditional TCP congestion control protocols might degrade significantly in some cases. One such feature is TCP Incast (Wu et al. 2013): when a client sends data request to many servers simultaneously and these servers return data to the client at the same time. In such a case, the buffer of switches will overflow, which will greatly degrade the throughput of the network. Researchers propose several techniques to address the TCP Incast problem. Another hot research issue in TCP for DCNs is multipath TCP (MPTCP). In topologies like Fat-Tree and BCube, there are multiple paths between a pair of hosts. In many cases, MPTCP can significantly increase the data throughput in DCNs.

Processing Platforms

Cloud computing (Armbrust et al. 2010) provides powerful computation capability to process the big data stored in data centers. According to the provided services, cloud computing can be classified into three categories:

- *Infrastructure as a Service (IaaS)*: In this type of service model, the cloud provides a set of highly scalable and automated computer resources to users in the form of bare hardware. The users can install operating system, deploy applications, or develop programmes by themselves. Examples of IaaS include Amazon AWS and Rackspace. The users can exploit the high scalability of IaaS, which allows them to dynamically scale up resources on demand instead of buying hardware outright.
- *Platform as a Service (PaaS)*: In this type of service model, the cloud provides a platform with components containing certain software for some applications. PaaS allows the customers to develop, run, and manage their applications on the platform, without building and managing low-layer infrastructure issues as in IaaS. PaaS provides a framework on which the users can create their customized applications. Examples of PaaS include Google App Engine and Windows Azure.
- *Software as a Service (SaaS)*: In this type of service model, the cloud delivers applications to users through Internet. In the SaaS model, the user needs not to download and install applications on their individual computers. The customer can focus on maintenance and support of their application, while the detailed technical issues such as data, servers, storage, and middleware are all handled by the software vendors. Examples of SaaS include Google Apps, Dropbox, and Salesforce.

Big Data Processing Framework

MapReduce (Dean and Ghemawat 2008) is a programming model for processing parallelizable

problems across huge volume of data using a cluster of computers. MapReduce can process both unstructured data stored in a filesystem or structured data stored in a database. It can take advantage of the locality of data, processing the data near the place where the data is stored in order to minimize communication overhead. When using MapReduce, the user specifies a map function and a reduce function. The map function takes a key/value pair as input and generates a set of intermediate key/value pairs. The reduce function merges the intermediate values with the same intermediate key.

A MapReduce program usually contains three steps:

- *Map*: In the map step, each worker node applies the map function to its local data and writes the output of the map function to a temporary storage. There is a master node that is in charge of partitioning data and send data to different worker nodes for processing, ensuring that only one copy of redundant input data is processed.
- *Shuffle*: In the shuffle step, the worker nodes redistribute data based on the output keys, such that all data belonging to one key are located on the same worker node.
- *Reduce*: In the reduce step, the worker nodes process each group of output data and merge the intermediate results in parallel.

There are two limitations of MapReduce. First, it cannot handle iterative algorithms, but a large number of data processing algorithms (e.g., many machine learning algorithms) are iterative. Second, its performance is greatly affected by the I/O operations. To overcome the limitations of MapReduce, the Spark (Zaharia et al. 2010) framework is developed. Compared with MapReduce, Spark significantly improves efficiency by doing in-memory computation. It provides caches to support iterative computation and multiple data sharing, which reduces cost in data I/O. It decreases the cost in writing intermediate result to HDFS by using the directed acyclic graph (DAG) model and reduces

redundant sorting and I/O operations by using multi-thread pool. Spark supports more flexible APIs than MapReduce and supports more languages.

Machine Learning for Big Data Processing

Machine learning are a powerful technique to extract useful information or knowledge from big data. For example, face recognition based on deep learning has been successful in many security-related applications in recent years. Machine learning-based big data processing usually requires a large volume of labelled training data, with which we can build very powerful models that sometime even outperform human being in some special tasks. Recently, how to design suitable network structure for machine learning models that run on a large number of computing nodes have attracted researcher's attention. This includes designing optimized network structure to reduce parameter transmission in learning models and adaptive learning model execution framework to reduce execution latency of the learning model.

Big Data Network Requirements

Different from traditional data processing, big data processing faces many challenges. First, besides structured data, in big data applications, most data are only partially structured or even unstructured. Second, the volume is extremely large, in the magnitude of PB or even EB. Third, the relationship among data is complex and unknown, making it difficult to use existing approaches to extract useful information from the data. As pointed out in Al-kahtani (2016), the challenges pose some requirements on big data networks, some of which are listed as below.

- *Network resiliency*: The availability of a network is critical to communication of distributed resources. Path diversity and failover can help improve the resiliency of the networks to failures. Path diversity means to utilize different routes rather than a single fixed route between a pair of communicating nodes. Failover means to have the ability to recognize failures and switch over to an alternate path quickly.
- *Network congestion mitigation*: When many big data applications launch at the same time, their data flow might work in a burst mode and might cause problems like Incast. This might cause congestion problems and consequently increase response latency. As previously described, network congestion can be mitigated by designing network topology and optimizing over TCP.
- *Network performance consistency*: Because the processing time in big data applications is usually dominated by the computational time rather than data transmission delay, big data networks are not affected a lot by network latency. For big data applications, one more important issue is to maintain high synchronicity, which refers to that different jobs in big data applications need to be executed in parallel in order to assist in accurate analysis. In such cases, any large decrease in network performance may trigger failures in the outcome.
- *Network future scalability*: As more and more big data applications will be ported to data centers, it is important for the data center architecture to have the ability of scale-up adaptively according to the requirement of applications.
- *Application awareness*: Big data network architecture typically use computer cluster environments such that the nodes can build large data sets. Different applications might pose different requirements on the performance of the data center. For example, some applications might need high bandwidth, but others might focus on low latency. If the network aims to support different tenants' applications with different requirements, it needs to support, differentiate, separate, and process various workloads with different requirements. When designing a big data network, the designer needs to consider all factors including coexistence issues of data flows, processes, and application resources.

Key Applications

Big data networks have been utilized in various domains to process huge volume of data. Some applications that utilize big data networks are listed below.

- *Key protein complexes identification:* How to identify the key protein complexes and their function models from a protein-protein interaction (PPI) network are most computation-intensive problems in the bioinformatics research (Stelzl et al. 2005). The high time complexity of traditional algorithms make them unsuitable for large-scale PPI networks. Currently, MapReduce and Spark have been used to process large-scale PPI networks (You et al. 2014; Harnie et al. 2017).
- *Recommendation systems in E-Business:* In E-commerce companies such as Amazon and Alibaba, how to predict the preferences of different customers and make correct recommendation for them is a key issue. Amazon makes recommendation decisions with big data, which significantly increases its profit. Amazon uses *item-to-item* collaborative filtering to obtain accurate recommendation results, which relies big data network to handle billions of customers and products.
- *Community discovery in social networks:* Community detection and discovery in large-scale social networks is another typical application that relies on big data network to process huge datasets. Large social media networks usually have a huge number of users, e.g., Facebook has more than 2 billion daily active users by 2017. In order to efficiently detect and discover communities in such large-scale social networks, MapReduce has been used (Guo et al. 2015).

Cross-References

- [Cloud Computing](#)
- [Data Center Network Virtualization](#)

References

- Al-Fares M, Radhakrishnan S, Raghavan B, Huang N, Vahdat A (2010) Hedera: dynamic flow scheduling for data center networks. In: Proceeding of NSDI, pp 89–92
- Al-kahtani MS (2016) Big data networking: requirements, architecture and issues. *Int J Wirel Mobile Netw* 8(6):1–15
- Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I et al (2010) A view of cloud computing. *Commun ACM* 53(4):50–58
- Bari MF, Boutaba R, Esteves R, Granville LZ, Podlesny M, Rabbani MG, Zhang Q, Zhani MF (2013) Data center network virtualization: a survey. *IEEE Commun Surv Tutor* 15(2):909–928
- Buyya R, Yeo CS, Venugopal S (2008) Market-oriented cloud computing: vision, hype, and reality for delivering it services as computing utilities. In: 10th IEEE international conference on high performance computing and communications, 2008 (HPCC'08). Ieee, pp 5–13
- Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: a distributed storage system for structured data. *ACM Trans Comput Syst (TOCS)* 26(2):4
- Chervenak A, Foster I, Kesselman C, Salisbury C, Tuecke S (2000) The data grid: towards an architecture for the distributed management and analysis of large scientific datasets. *J Netw Comput Appl* 23(3):187–200
- Dean J, Ghemawat S (2008) Mapreduce: simplified data processing on large clusters. *Commun ACM* 51(1):107–113
- Dowd K, Severance CR, Loukides M (1998) High performance computing-RISC architectures, optimization and benchmarks. O'Reilly, Paris
- Gantz J, Reinsel D (2012) The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze Future 2007(2012):1–16
- Gligor VD, Shattuck SH (1980) On deadlock detection in distributed systems. *IEEE Trans Softw Eng* 6(5): 435–440
- Guo K, Guo W, Chen Y, Qiu Q, Zhang Q (2015) Community discovery by propagating local and global information based on the mapreduce model. *Inf Sci* 323:73–93
- Harnie D, Saey M, Vapirev AE, Wegner JK, Gedich A, Steijaert M, Ceulemans H, Wuyts R, De Meuter W (2017) Scaling machine learning for target prediction in drug discovery using apache spark. *Future Gener Comput Syst* 67:409–417
- Nie F, Zhu W, Li X (2017) Unsupervised large graph embedding. In: AAAI, pp 2422–2428
- Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122(6):957–968

- Strogatz SH (2001) Exploring complex networks. *Nature* 410(6825):268
- Sun Y, Song H, Jara AJ, Bie R (2016) Internet of things and big data analytics for smart and connected communities. *IEEE Access* 4: 766–773
- Wu H, Feng Z, Guo C, Zhang Y (2013) Ictcp: incast congestion control for tcp in data-center networks. *IEEE/ACM Trans Netw (ToN)* 21(2): 345–358
- You Z-H, Yu J-Z, Zhu L, Li S, Wen Z-K (2014) A mapreduce based parallel svm for large-scale predicting protein–protein interactions. *Neurocomputing* 145:37–43
- Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: cluster computing with working sets. *HotCloud* 10(10-10):95

Big Network Data

Zichuan Xu¹, Qiufen Xia², and Lin Yao²

¹School of Software, Dalian University of Technology, Jinzhou Area, Dalian, China

²International School of Information Science and Engineering, Dalian University of Technology, Jinzhou Area, Dalian, China

Synonyms

[Big data in networks](#); [Network big data](#)

Definitions

Big network data refers to big data that is generated or collected in different networks varying from social networks, communication networks, transportation networks, the World Wide Web (WWW), biological networks, citation networks, etc., that is, datasets in networks that are so voluminous and complex that traditional data processing methods are inadequate to deal with them, while advanced data analytics methods are needed to extract value from them. The five dimensions to big data known as Volume, Variety, Velocity, Veracity, and Value still apply to big network data.

Historical Background

The evolution of big network data is in line with the evolving of large-scale distributed systems (Akyildiz et al. 2002; Armbrust et al. 2009; Fernando et al. 2013), the Internet, and networking technologies, which turned the processing of big data and obtaining valuable information from the big data increasingly challenging and difficult. Namely, the scale and complexity of network data begin increasing with the invention of computer networks designed to exchange data in the 1950s. From the establishment of Wide Area Networks (WAN) in 1960s, the wide adoption of TCP/IP in 1970s to the most recent development in 5G communication networks, volumes, types, and generation speeds of network data have been increasing dramatically.

One of the first milestones in the history of network data was the arrival of Salesforce.com in 1999, which is a pioneer by delivering enterprise applications via a simple Internet website hosted in servers. This allows the data being generated in the servers with rationale databases. The next development of network data was Amazon Web Services in 2002, which provided a series of cloud-based services including storage and computation. Meanwhile, by 2006, Facebook and Twitter both became available to users throughout the world, which accelerates the generation of not only user data but also large-scale graph data that characterize user relations. With the increasing number and advancement of such cloud services and online social networks, more and more enterprise, governments, and individual users are relying on cloud services for either high efficiency for their daily businesses or low cost for expenditure reduction. With the continuous running of such cloud services, large volumes of data generated at unprecedented rates from sensors, emails, Internet transactions, click streams, user preference of purchases, etc. are getting attention of many enterprises as it poses valuable patterns and business insights.

In the past decade, the availability and growing expectation of ubiquitous connectivity is one of the most transformative technology driving big network data. Whether it is for checking

email, carrying a voice conversation, or browsing web pages, mobile users can access these online services regardless of location, time, or circumstance: at the office, on a subway, while in flight, etc. Such ubiquitous mobile accesses to services allow the users to generate data anywhere any-time at a unprecedented rate.

Most recently, the techniques of 5G, Software-Defined Networking (SDN) and Network Function Virtualization are pushing the big network data into a new age. On one hand, these techniques further enlarges the capacity of networks by allowing more users, thereby future contributing to the volumes and diversity of big network data. On the other hand, they enhance the efficiency of big data analytics by allowing more efficient data transfers in networks. For example, SDN that separates the network control logic (the control plane) from the data forwarding plane is a new promising technology to enable efficient data transfers via the flexible management of network resources.

Foundations

Big network data involves the data that is related to any networking technologies, varying from communication networks to social networks. According to types of networks, big network data usually can be classified into the following categories:

- **Big data in wireless networks:** Given the myriad different use cases and applications of wireless networks, there exist dozens of wireless networking technologies, each of which has its own performance characteristics and optimized for a specific task and context. For example, there are over a dozen of widespread wireless technologies in use: WiFi, Bluetooth, ZigBee, NFC, WiMAX, LTE, HSPA, EV-DO, 3G standards, satellite services, etc. Big network data in different wireless networks is diverse in terms of collecting, storing, and processing methods, as they are designed for different usage scenarios and handle user data in different ways. For example, wireless sensor networks focus on the collection of multi-dimensional (location, time) sensing data in an energy-efficient and bandwidth-saving manner (Akyildiz et al. 2002; Cardei et al. 2002; Liang 2006), and thus various data collection, data compression, data transmission, and coding methods are proposed. Internet-of-Things (IoT) networks are another major source of big network data. Given that sensors are used in nearly all industries, the IoT is expected to produce a huge amount of data (Ahmed et al. 2017). The volume and diversity depend on specific application of the IoT.
- **Big data in cloud networks:** With an exponential increase in cloud services for smart device users, there is an increase in the bulk amount of data generated by the services. Generally, most service providers, including Google, Amazon, Microsoft, and so on, have deployed a large number of geographically distributed data centers that are interconnected by wide-area networks – known as *cloud networks*, to process huge amounts of big network data so that users can get quick response. There are two directions related to big data processing in such cloud networks. One direction is the design of efficient parallel processing frameworks and tools to achieve real-time process of big data within one data center. Hadoop and SPARK are examples for this purpose, which are widely used by cloud service providers (Singh and Reddy 2014; Reyes-Ortiz et al. 2015; Zaharia et al. 2016). Another direction is the design of novel network architectures, communication protocols, routing algorithms, etc. to alleviate the bottleneck of underlying network infrastructures in terms of transfer congestions due to big data transfers (Gu et al. 2014; Jiao et al. 2014; Xia et al. 2016, 2017a,b).
- **Big graph data:** Social networking sites (e.g., Facebook, Twitter, and LinkedIn) keep expanding in terms of the online active users and the data generated by the

users. The data in such social networking sites can usually be modeled as graphs, with their nodes representing users and edges representing relations. Such graph data are being generated at an unprecedented rate; they are now measured in terabytes and heading toward petabytes, with more than billions of nodes and edges. For example, it is reported that WWW now contains more than 50 billion web pages and more than one trillion URLs (WWW 2018). A recent snapshot of the friendship network Facebook contains 800 million nodes and over 100 billion links (Facebook 2018). One of the most fundamental problems related to such big graph data is finding cohesive groups that can be regarded as *communities* such as friends, co-workers, neighbors, etc. (Zhou et al. 2009). For example, in web graphs, they are likely groups of web pages about the same or related topics. Another fundamental problem is to identify highly “influential” communities and individuals from such large networks as they govern the entire network behaviors and thus bring enormous social and economic benefits (Ding et al. 2007).

- **Big monitoring data:** The complexity of large-scale networks is constantly increasing. With more services being offered and the continuous growth of bandwidth-hungry video-streaming services, network and server infrastructures are becoming extremely difficult to monitor, since the volume and complexity of monitoring data are increasing in an unprecedented rate. Network Traffic Monitoring and Analysis (NTMA) is to process big, heterogeneous, and high-speed data related to various status of the networks (Bär et al. 2014). Although big monitoring data preserve all the “V”s of big data, the traditional techniques for general purpose big data cannot be directly applied to NTMA. The reason behind is that traditional big data technologies lack efficient and effective methods for real-time traffic analysis (Bär et al. 2014; Liu et al. 2014).

Key Applications

Efficient and effective analysis of big network data has wide applications in many areas spanning from Internet-of-Things (IoT) industrial services to every economic aspect of people’s daily lives. The key applications of big network data and its corresponding analytic techniques are summarized as follows, from the perspectives of Internet of Things, communication and cloud services, trading promotion and decision-making, and innovative networking technologies.

- *Internet of Things:* The efficient and continuous processing of big network data in IoTs can contribute to the correct operation and development of IoT applications. For example, data collection efficiency in IoTs can be promoted dramatically if analytic results on big network data show a smart distribution of sensor locations where important values are obtained. In addition, to generate benefits from IoT, enterprises need analytic techniques for big network data, via which they can collect, manage, and analyze a massive volume of data from sensors in a scalable and cost-effective manner. In this context, big network data can assist in consuming, reading, and analyzing diverse enterprise data on information that is collected by enterprises.
- *Network and cloud services:* Big network data is a significant part of network and cloud services. The efficient analysis of big network data in return can contribute to the advancement of such services. For example, the users’ preference of cloud services can be obtained through analyzing user access information, thereby contributing new cloud services. From the views of network service providers, by analyzing the locations of their mobile users based on big network data, strategic deployments of their infrastructures can be performed to maximize the revenue of network service providers.
- *Trading promotion and decision-making:* Identifying highly influential communities

and individuals based on big graph data in large-scale networks is significant in terms of social and economic benefits. For example, identifying highly influential communities in a large collaboration network can help to find world-leading research teams and potential research directions. Also, identifying highly influential communities in trading and economic networks will enable enterprises to better promote products and enhance business opportunities. On the other hand, identifying most influential individuals is significant to disparate applications, including accelerating information propagation, controlling rumors and diseases, designing search engines, and understanding hierarchical organization of social and biological networks.

- Innovative networking technologies: Big monitoring data in networks characterize the running status of a network that is supported by a specific networking technology. For example, by consistently monitoring and analyzing security attacks in networks, novel security protocols can be designed to prevent future security breaches. Also, analytic results based on the monitoring data related to network resource allocations will have a major impact on the design of network resource allocation strategies and mechanisms.

Cross-References

- Big Data Networks
- Internet of Things
- Ubiquitous Big Data
- Wireless Big Data

References

- Ahmed E, Yaqoob I, Hashem IAT, Khan I, Ahmed AIA, Imran M, Vasilakos AV (2017) The role of big data analytics in internet of things. *Comput Netw* 129:459–471. <https://doi.org/10.1016/j.comnet.2017.06.013>, <http://www.sciencedirect.com/science/article/pii/S1389128617302591>. Special Issue on 5G Wireless Networks for IoT and Body Sensors
- Akyildiz I, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. *Comput Netw* 38(4):393–422. [https://doi.org/10.1016/S1389-1286\(01\)00302-4](https://doi.org/10.1016/S1389-1286(01)00302-4), <http://www.sciencedirect.com/science/article/pii/S1389128601003024>
- Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Zaharia M (2009) Above the clouds: a berkeley view of cloud computing. Technical report
- Bär A, Finamore A, Casas P, Golab L, Mellia M (2014) Large-scale network traffic monitoring with dbstream, a system for rolling big data analysis. In: 2014 IEEE international conference on big data (Big Data), pp 165–170. <https://doi.org/10.1109/BigData.2014.7004227>
- Cardei M, MacCallum D, Cheng MX, Min M, Jia X, Li D, Du DZ (2002) Wireless sensor networks with energy efficient organization. *J Interconnection Netw* 03(03n04):213–229. <https://doi.org/10.1142/S021926590200063X>
- Ding B, Yu JX, Wang S, Qin L, Zhang X, Lin X (2007) Finding top-k min-cost connected trees in databases. In: 2007 IEEE 23rd international conference on data engineering, pp 836–845. <https://doi.org/10.1109/ICDE.2007.367929>
- Facebook (2018) Facebook. <http://www.facebook.com/press/info.php?statistics>
- Fernando N, Loke SW, Rahayu W (2013) Mobile cloud computing: a survey. *Futur Gener Comput Syst* 29(1):84–106. <https://doi.org/10.1016/j.future.2012.05.023>, <http://www.sciencedirect.com/science/article/pii/S0167739X12001318>. Including Special section: AIRCC-NetCoM 2009 and Special section: Clouds and Service-Oriented Architectures
- Gu L, Zeng D, Li P, Guo S (2014) Cost minimization for big data processing in geo-distributed data centers. *IEEE Trans Emerg Top Comput* 2(3):314–323. <https://doi.org/10.1109/TETC.2014.2310456>
- Jiao L, Lit J, Du W, Fu X (2014) Multi-objective data placement for multi-cloud socially aware services. In: IEEE INFOCOM 2014 – IEEE conference on computer communications, pp 28–36. <https://doi.org/10.1109/INFOCOM.2014.6847921>
- Liang W (2006) Approximate minimum-energy multicasting in wireless ad hoc networks. *IEEE Trans Mobile Comput* 5(4):377–387. <https://doi.org/10.1109/TMC.2006.1599406>
- Liu J, Liu F, Ansari N (2014) Monitoring and analyzing big traffic data of a large-scale cellular network with hadoop. *IEEE Netw* 28(4):32–39. <https://doi.org/10.1109/MNET.2014.6863129>
- Reyes-Ortiz JL, Oneto L, Anguita D (2015) Big data analytics in the cloud: spark on hadoop vs mpi/openmp on beowulf. *Proc Comput Sci* 53:121–130. <https://doi.org/10.1016/j.procs.2015.07.286>, <http://www.sciencedirect.com/science/article/pii/S1877050915017895>. iNNS Conference on Big Data 2015 Program, San Francisco, 8–10 Aug 2015
- Singh D, Reddy CK (2014) A survey on platforms for big data analytics. *J Big Data* 2(1):8. <https://doi.org/10.1186/s40537-014-0008-6>
- WWW (2018) Wwww. <http://www.worldwidewebsite.com/>

- Xia Q, Xu Z, Liang W, Zomaya AY (2016) Collaboration- and fairness-aware big data management in distributed clouds. *IEEE Trans Parallel Distrib Syst* 27(7):1941–1953. <https://doi.org/10.1109/TPDS.2015.2473174>
- Xia Q, Liang W, Xu Z (2017a) Data locality-aware big data query evaluation in distributed clouds. *Comput J* 60(6):791–809. <https://doi.org/10.1093/comjnl/bxw101>
- Xia Q, Liang W, Xu Z (2017b) The operational cost minimization in distributed clouds via community-aware user data placements of social networks. *Comput Netw* 112:263–278. <https://doi.org/10.1016/j.comnet.2016.11.012>
- Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ, Ghodsi A, Gonzalez J, Shenker S, Stoica I (2016) Apache spark: a unified engine for big data processing. *Commun ACM* 59(11):56–65. <https://doi.org/10.1145/2934664>
- Zhou Y, Cheng H, Yu JX (2009) Graph clustering based on structural/attribute similarities. *Proc VLDB Endow* 2(1):718–729. <https://doi.org/10.14778/1687627.1687709>

Binding

- [Proxy Mobile IPv6](#)

Biomedical Internet of Things

- [Internet of Medical Things](#)

Blockchain

Yuan Zhang
University of Electronic Science and Technology
of China, Chengdu, China

Synonyms

[Blockchain](#); [Blockchain-based currencies](#); [Proofs of stake](#); [Proofs of work](#)

Definition

A blockchain is a linear collection of data elements called *block*, where all blocks are linked to form a chain and secured using cryptography, and newly generated blocks are continuously chained to the blockchain in an untrusted environment. To date, there is still lack of formal definitions on the blockchain that can be accepted by both the academia and industry.

Historical Background

The Blockchain technique is derived from the Bitcoin which is first proposed by Nakamoto in 2008. The primary objective of Bitcoin is to propose a secure payment system without a trusted party such that online payments can be sent directly from one user to others without going through a financial institution. This is achieved by a distributed ledger which accounts for the ownership of coins. The key challenge is to resist double spending attacks, where an adversary could issue two transactions in parallel so as to transfer the same coin to different recipients. Essentially, enabling the distributed ledger to be secure against double spending attacks boils down to the Byzantine Generals problem (BGP). Existing solutions to solve BGP mainly focus on the permission model, where the participants in the system are predetermined (or participants know each other). However, the participants who maintain the ledger in a secure digital payment system are under a permissionless model, where anyone can join or leave the system without permitted by a centralized or distributed authority.

Bitcoin first achieved consensus on the distributed ledger in the permissionless model, where it cleverly integrates existing primitives and solutions from decades of research into one system to solve the fundamental problems existing in digital currencies in a practically viable way. Bitcoin is also the first cryptocurrency that is widely applied in large-scale networks. Due to the success of Bitcoin, its underlying technique called “blockchain” is extracted,

which is of independent interest. Intuitively, the blockchain is a technique that maintains a public, immutable, and ordered ledger of records among all participants. It enables the participants to securely and periodically add new records into the ledger without trusting each other. It also guarantees that the records that have been added into the ledger cannot be removed or modified such that all honest participants have a consistent view of the ledger. More recently, the blockchain technique has been further investigated and one of the most prominent manifestations is to enable smart contracts. As a disruptive technique with profound implications, blockchain is transforming the very nature of how users including individuals and enterprises conduct transactions and perform smart contracts.

Generally, the blockchain technique can be classified into two types: public blockchain and private blockchain. For a public blockchain, the participants who maintain the blockchain are subject to the permissionless model, which means that anyone can join or leave the blockchain system without getting permission from a centralized or distributed authority. The most prominent manifestation of public blockchain is blockchain-based currencies, such as Bitcoin and Ethereum. The private blockchain (including the consortium blockchain) is derived from its public counterpart, where the participants who maintain the blockchain are authorized by blockchain managers or the managers themselves. At this point, the key difference between the private blockchain and distributed backup systems is that once a record is added to the blockchain, as long as the majority of participants is accessible to the adversaries, the record cannot be removed or modified, even if the blockchain manager becomes the adversary.

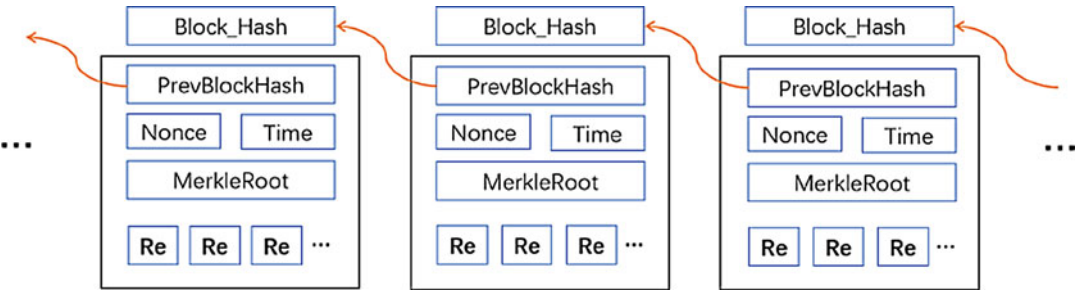
Foundations

Most of existing blockchain systems are either based on the proofs of work (PoW) or based on the proofs of stake (PoS) (Kiayias et al. 2017; Pass and Shi 2017; Tschorsch and Scheuermann 2016; Underwood 2016; Wood 2014). Both PoW

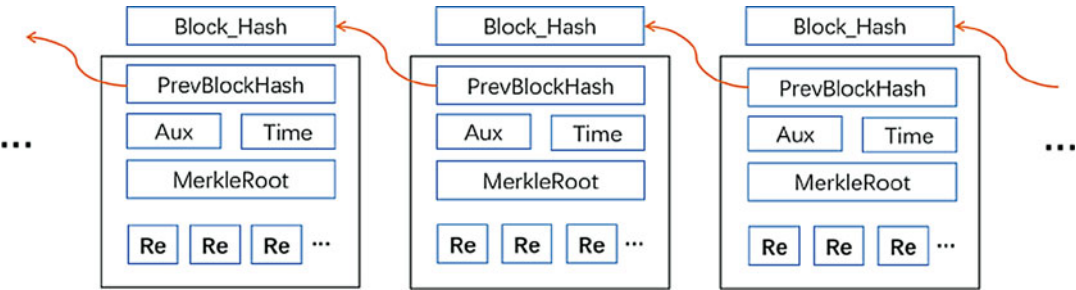
and PoS are key components to implement a consensus algorithm such that newly generated records can be securely added to the blockchain. PoW is a cryptographic primitive that enables a prover to demonstrate to a verifier that she/he has performed a certain amount of computational work in a specified interval of time. In the PoW-based blockchain, PoW is used to build a bridge between a user's identity and her/his hashrate, such that the more hashrate in the blockchain network, the higher probability that the content of the new block is determined by the participant. A typical PoW-based blockchain has the form shown in Fig. 1, where Block Hash denotes the hash value of current block, PreBlockHash denotes the hash value of the last block, Nonce denotes the PoW solution, Time denotes the timestamp of current block, Re denotes the record stored in the blockchain, and MerkleRoot denotes the root value of a Merkle hash tree formed by all records recorded in current block.

In the PoW-based blockchain, the participants need to solve the PoW puzzle to fight for generating the new block; this results in striking energy demands. Actually, the PoW-based blockchain has even been labeled an "environmental disaster" (Dziembowski et al. 2015). As such, a blockchain system that is constructed on another mechanism, rather than PoW, to obviate the huge energy requirements is urgently needed. Essentially, PoW in blockchain facilitates a practical randomized "leader election" process that elects one of the participants to issue the next block, where the probability to be the leader for a participant is proportional to her/his hashrate.

A natural alternative mechanism to replace PoW is the notion of "proof of stake" (PoS). In PoS-based blockchain, the probability to be the leader for a participant is proportional to her/his stake according to the current blockchain ledger. A basic observation to use such mechanism is that if the higher ratio of shareholding one has, the stronger motivation to retain the system security she/he has. A key challenge to design a PoS-based blockchain is to simulate the randomized leader election process. In most of existing PoS-based blockchain systems, this is achieved by a cryptographic protocol that participants jointly



Blockchain, Fig. 1 A typical PoW-based blockchain



Blockchain, Fig. 2 A typical PoS-based blockchain

generate a random number to elect the leader. A typical PoW-based blockchain has the form shown in Fig. 2, where Aux denotes the auxiliary information on the leader election process.

Except for PoS, some other mechanisms have been proposed to replace PoW, for example, proofs of space. In the proofs of space setting, a prover demonstrates to a verifier that she/he has significant amounts of disk space.

Applications in Wireless Networks

Blockchain-based identity management for wireless networks. Blockchain can be utilized as a database to record all users’ identities in wireless networks without a trusted authority (Garman et al. 2014). Since the blockchain is inherently resistant to forgery and modification, after the users’ identities are recorded into the blockchain, anyone cannot tamper with the identities. Compared with existing systems, the blockchain-based identity management

system does not require a trusted authority and thereby can resist the single-point-of-failure problem.

Blockchain-based random number generator for wireless networks. In wireless networks, random numbers play an important role to construct systems and protocols. In a PoW-based blockchain system, the hash value (Block Hash in Fig. 1) of each block on the chain only can be determined after a valid nonce is computed and the block is chained to the blockchain. Due to the inherent resistance against modification of blockchain, once the block is chained to the blockchain, its hash value is deterministic and would never be changed. As such, given a time t , if t is a past time, the hash value of the latest block that has appeared since t in the PoW-based blockchain is deterministic and can be extracted efficiently; if t is a future time, the hash value is unpredictable (Zhang et al. 2015, 2016). Therefore, the PoW-based blockchain can be utilized to construct a time-dependent random number generator.

Blockchain-based data sharing for wireless networks. In wireless networks, multiple network systems need to share data with each other. However, due to the differences on the data form, software version, network delay, or even node location, sharing data is always a challenging problem for different wireless network systems (Kosba et al. 2016). Since the blockchain is publicly accessible to all participants, it enables the data sharing among different systems to be feasible and efficient. Multiple systems can joint maintain a blockchain to record the data to be shared, which build bridges between these systems to share data.

Cross-References

- [Blockchain: Enabling Future Internet with Intrinsic Security](#)
- [Distributed Storage Security](#)

References

- Dziembowski S, Faust S, Kolmogorov V, Pietrzak K (2015) Proofs of space. In: Proceedings of the CRYPTO. Springer, Santa Barbara, CA, pp 585–605
- Garman C, Green M, Miers I (2014) Decentralized anonymous credentials. In: Proceedings of the NDSS. San Diego, CA, pp 23–26
- Kiayias A, Russell A, David B, Oliynykov R (2017) Ouroboros: a provably secure proof-of-stake blockchain protocol. In: Proceedings of the CRYPTO. Springer, Santa Barbara, CA, pp 357–388
- Kosba A, Miller A, Shi E, Wen Z, Papamanthou C (2016) Hawk: the blockchain model of cryptography and privacy-preserving smart contracts. In: Proceedings of the SP. IEEE, San Jose, CA, pp 839–858
- Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. Technique Report, <http://wfc-knowledgecentre.com/wp-content/uploads/2016/07/Bitcoin-A-Peer-to-Peer-electronic-Cash-System.pdf>
- Pass R, Shi E (2017) Fruitchains: a fair blockchain. In: Proceedings of the PODC. ACM, Washington, DC, pp 315–324
- Tschorsch F, Scheuermann B (2016) Bitcoin and beyond: a technical survey on decentralized digital currencies. IEEE Commun Surv Tutor 18(3):2084–2123
- Underwood S (2016) Blockchain beyond Bitcoin. Commun ACM 59(11):15–17
- Wood G (2014) Ethereum: a secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper, pp 1–32
- Zhang Y, Xu C, Yu S, Li H, Zhang X (2015) SCLPV: secure certificateless public verification for cloud-based cyber-physical-social systems against malicious auditors. IEEE Trans Comput Soc Syst 2(4):159–170
- Zhang Y, Xu C, Li H, Liang X (2016) Cryptographic public verification of data integrity for cloud storage systems. IEEE Cloud Comput 3(5):44–52

Blockchain: Enabling Future Internet with Intrinsic Security

Kai Wang¹ and Dongchao Guo²

¹Harbin Institute of Technology, Weihai, P. R. China

²Beijing Information Science and Technology University, Beijing, P. R. China

Synonyms

[Blockchain](#); [Future Internet architecture](#); [Network security](#)

Definitions

The term, “blockchain,” has various meanings in different context. It is often referred to as the set of the technologies constituting the blockchain-based system, including but not limited to the ledger, the consensus algorithm, the hash algorithm, the asymmetric cryptography, and so on (Narayanan et al. 2016). In the context of data structure, the term “blockchain” is referred to the chain of blocks. Each block includes a bundle of transactions bound together and is secured via cryptography. Each transaction is considered as connecting two entities which are identified by the cryptographic addresses generated from the public keys of them. The contents in the transaction indicate the transfer of valued data between

the involved entities. The valued data contained in each transaction can be cryptocurrency, business records, finance incoming, certificate of a care, smart contract, or anything in the digital form that is valuable. For instance, the transaction may indicate the change of the ownership of some amount of cryptographic coins from one person to another. Many transactions are combined into a block, and then this block can be added as the latest block of the chain after it is eventually validated via the kind of process named “consensus” following the rules of some certain consensus protocol. Following similar process, the subsequent blocks are appended one after another to form a growing chain, actually the blockchain, to store all the transactions that have been validated and witnessed by every single participant in the blockchain system.

The Emergence of Blockchain Technology

The relentless pursuit of the better way of payment, which is secure, anonymous, and trusted, is driven by the passion of the cypherpunk (Hughes 1997), where people advocate the spread of the usage of cryptographic technologies to the public. It is widely accepted that the first generation blockchain technology is exemplified by the decentralized cryptocurrency, among which the Bitcoin, proposed by Satoshi Nakamoto in 2008 (Nakamoto 2008), may be the first and most famous one. As the fast growth of the cryptocurrency market capitalization, the underlying technology, blockchain, is soon widely recognized by the whole world. The Ethereum, proposed by V. Buterin, G. Wood, and J. Lubin in 2014 (Wood 2014), is believed to start the era of blockchain 2.0, for it incorporates a Turing-complete language (i.e., the so-called smart contract) making it possible to implement all kinds of decentralized applications. The blockchain-based system, such as Bitcoin and Ethereum, featuring every peer freely joining in and fairly competing to add new records to the chain, is said to be “public.” In contrast, the system, like Hyperledger, where only some super recorders have the right to do

that, is regarded as the private blockchain or the consortium blockchain.

These days, blockchain supports not only the transfer of cryptocurrency in a decentralized and trusted way but also enables a wide variety of decentralized applications. It has great technological, economical, and even social significance. Everyone, to some extent, is enabled to the right to issuing private moneys via the blockchain technology, and the private money has its economic origins from the denationalization of money proposed by Friedrich Hayek in 1970s (Hayek 1990). Since the economy these days is based on the fiat, the blockchain-based cryptocurrency has the potential to change the world significantly. More importantly, it may enable new types of economic organizations and governance via smart contracts to enable the decentralized worldwide cooperation.

On the other kind, it may not only help to deal with the challenges raised by the fast development of mobile Internet and Internet of Things (IoT) but also contribute to the establishment of the future information infrastructure. Actually, some efforts have been paid into this. For instance, the MIT Connection Science Research Initiative aims at building up a future Internet architecture of trusted data (Pentland et al. 2016). In addition, Blockstack is proposed to construct a blockchain-based decentralized DNS featuring low searching latency (Ali et al. 2016).

There are also some challenges in the development of the blockchain technology and its applications that we need to deal with (Vukolić 2015; Zheng et al. 2018). The public blockchain trilemma is the trade-off between the performance, the decentralization, and the security in a blockchain-based system. At its early state, the average number of legal transactions added to the chain, proved via the Proof-of-Work (PoW) consensus protocol, is merely less than 100, while keeping good decentralization. At the cost of decentralization as well as security, the BFT protocol employed by the consortium blockchain may increase the throughput 1000 times more than the PoW protocol used by the public blockchain. The Algorand (Gilad et al. 2017), started by the

Turing award winner Silvio Micali, could be considered as a hybrid of the public blockchain and the consortium blockchain and may keep a better balance, although it is much more difficult to implement and a bit far from being used in real systems. Theoretically, the adversary may take control of the blockchain system if it holds more than one half computation resources of the PoW-based blockchain or more than one third of the BFT-based blockchain. However, in practice the computation resources that need to hold to control the blockchain system is far less than half, due to the emergence of large mining pools.

Some Critical Features of Blockchain

Some key features make it possible for the blockchain to be not only a promising ICT innovation but also an institutional technology (Aste et al. 2017; Davidson et al. 2016). Before elaborating the potential impact of the blockchain over our world, we would like to involve into the details of the blockchain itself and depict some critical features.

The blockchain-based system can be divided into the data layer, the consensus layer, the contract layer, and the application layer. The ledger is the most important data structure in the data layer. It is composed of ordered blocks which are in turn composed of validated transactions or smart contracts. The validity process, or in other words, the consensus process, is such that it ensures the chronological order in which transactions, smart contracts, and information have been created or executed, in a decentralized way. We can imagine the ordered chain of blocks as the chronicle of the history of the cyberspace, and the chronicle is composed by no means of an authority. Eliminating the central authority, every peer in the system is entitled to the access to the records in the chain. In other words, the records are traceable and transparent to every peer. As mentioned above, every peer competes to add new record, i.e., the block, to the chain, following the common consensus protocol. Once being validated and approved, the order could hardly be changed, unless the adversary takes control of big enough computation resources. In this

sense, the blockchain is considered as tamper-proof or immutable. The historic truth recorded in the chain has been validated by the whole community of peers. The smart contracts programmed based on the Turing-complete language enable the creation of decentralized applications and even the decentralized autonomous organizations which operate without human beings' involvement and without resorting to any third-party authority providing trust. To sum up, the key features of blockchain are the decentralization, the immutability, the self-execution, and the transparency.

Although one can enumerate a wide variety of areas where the blockchain technology could be applied to, we would like to elaborate the effect of the blockchain technology in a more general way. Through the public regulations coded intrinsically in the core protocol of the blockchain-based system which are transparent to all participants, it is possible to build trust between individuals and organizations in the cyberspace. Through the consensus process among the majority of the participants, it succeeds in establishing the regulations of cooperation so that the cooperation between untrusted organizations could continue. The transparency and the immutability are critical to eliminate the information asymmetry (Aboody and Lev 2000). As a result, the cost of trades and negotiations would be reduced dramatically. Also, the feature of transparency helps to break the silo of data storage in the current information management systems. It seems that the blockchain has the potential to endow the virtual cyberspace with some kind of reality and continuity. One could regard the records in the blockchain as the history of the cyberspace. Besides, some believe that blockchain is more like an innovation facilitating new types of organization and governance.

Toward the Future Internet Architecture with Built-In Security

Security Drawbacks of Current Internet Architecture

The Internet, as an information platform, has achieved a great success due to its convenience

in information dissemination. One of the key factors of this promising information platform is the efficiency and simplicity of the IP protocol in data transmission. On the other hand, the IP-based data delivery, as a kind of packet switching, has its disadvantages and drawbacks. For example, the authenticity and integrity of the transferred data can only be guaranteed by the connection that supports data transmission between the communication endpoints (e.g., a client requesting for data and a server responding with the requested data). The data can be trusted only when its provider that is connected with the requester is trusted during the data transmission. Whenever the connection is broken, the data cannot be trusted if it is retrieved from other places, even it is indeed the same with what the trusted provider supplies. In this sense, the trust is based on the connection in the current Internet. The non-repudiation of data cannot be protected by traditional Internet protocols.

Furthermore, the ownership of data in the Internet is difficult to protect after the data is uploaded into public cloud or acquired by some third-party enterprises or institutions. Whether the data is used for commercial purposes and how much business gain is achieved are not transparent to the original data owners. As a consequence, the data owners cannot benefit from any kind of data sharing although they produce and provide the data. The data abusers are not likely to get punished, which harms the data sharing in current Internet. Given the fact that the big data could be considered as the fuel for artificial intelligence, this shortcoming of the existing Internet is also detrimental to the development of advanced information technology.

Enabling Future Internet with Intrinsic Security

Fortunately, the blockchain technology is capable of enabling the current Internet with secure and trusted data storage, data access authorization, data computing, and data sharing between untrusted entities. It can provide trusted data ownership guarantee based on some consensus process between all involved Internet parties and can help virtual participants

to enable tamper-proof data or value transfer. As a result, it can help to achieve a real decentralized and secure data management in cyberspace.

Firstly, the blockchain technology can provide the consensus to claim the ownership of all scattered data stored in the Internet with cryptograph protect. If the data ownership is registered in the blockchain-based systems, it is hard for some certain third-party centralized entities to hide or abuse the data, as the data ownership can be publicly validated by all participants in any blockchain ecosystem. In the future Internet with blockchain protection on data ownership, the data of each Internet user can be stored distributed in the owner's devices or maintained by some neutral third-party data servers. Each data has its own identifier, e.g., the hash of the data. Each user also has an identifier, e.g., the public key or a bitcoin-like address of the owner of the data which is associated with the private key. The binding of the identifier of data and its corresponding owner (an Internet user) is recorded as a transaction which can only be controlled by the data owner. This binding relationship is validated by blockchain miners and then broadcast it to all the participants in this blockchain system, to reach the consensus on this data ownership. As a result, the abuse of the data is very difficult for any centralized entity since the public witness is achieved on the data ownership.

Secondly, the blockchain technology can achieve data access authorization with smart contracts, to support unified access control for different entities with different incentives. For different data with different formats or with different application functions, the blockchain provides a mechanism to generally quantify the access authorization by token incentives. Specifically, different data may have different values, which are measured with different types or number of tokens. The logic of access authorization on the data is implemented in the blockchain as smart contracts, which formulates the rules of data access and what kind of tokens and how much that requesters should pay for the data. No matter for which kind of data, the user that requests this data only need to trigger a proper smart contract on the blockchain and

transfer enough tokens to this smart contract to pay for this data. The tokens enable the unified access control for all the data in future Internet. If anyone wants to access any data, enough tokens are the only thing that should be guaranteed. That is, token is the connection bridge of every Internet user and data.

Finally, the blockchain technology can support tamper-proof records for transfer of data ownership by resorting to the immutable distributed ledgers. For every transaction in every block, the transfer of data ownership is witnessed by the public in the blockchain ledger, including the change of data ownership and the tokens that the data requester paid for. Furthermore, this transfer of data ownership is continually validated by the blockchain miners, as well as witnessed by all the participants across multiple autonomous domains. In this way, the tamper-proof records for the transfer of data ownership are confirmed in the immutable distributed ledgers by every single participant, making the transfer of data ownership trusted and verifiable.

Overall, the control of data ownership is a formidable challenge for current Internet technologies, yet the emerging blockchain technology gives a chance to solve this problem efficiently and effectively. Consequently, the blockchain technology may finally become one of the most promising candidates to achieve a future Internet with intrinsic security.

Cross-References

- [Access Control](#)
- [Authentication](#)
- [IoT Forensics](#)
- [Privacy-Preserving Data Collection](#)

References

Aboody D, Lev B (2000) Information asymmetry, R&D, and insider gains. *J Financ* 55(6):2747–2766

- Ali M, Nelson J, Shea R, Freedman M (2016) Blockstack: a global naming and storage system secured by Blockchains. In: Proceedings of USENIX annual technical conference (ATC). USENIX, Denver
- Aste T, Tasca P, Di Matteo T (2017) Blockchain technologies: the foreseeable impact on society and industry. *Computer* 50(9):18–28
- Blockstack. <https://blockstack.org/>
- Davidson S, De Filippi P, Potts J (2016) Economics of Blockchain. Social Science Electronic Publishing. <https://doi.org/10.2139/ssrn.2744751>, <https://ssrn.com/abstract=2744751>
- Gilad Y, Hemo R, Micali S, Vlachos G, Zeldovich N (2017) Algorand: scaling byzantine agreements for cryptocurrencies. In: Proceedings of 26th symposium on operating systems principles (SOSP). ACM, Shanghai
- Hayek FA (1990) Denationalisation of money. The argument refined: an analysis of the theory and practice of concurrent currencies. Institute of Economic Affairs, London
- Hughes E (1997) A cypherpunk's manifesto. The electronic privacy papers: Documents on the Battle for Privacy in the age of urveillance. ISBN 0-471-12297-1, John Wiley & Sons, Inc. New York, NY, USA, pp 285–287. <https://dl.acm.org/citation.cfm?id=285725>
- Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. Technique report. <http://wfc-knowledgecentre.com/wp-content/uploads/2016/07/Bitcoin-A-Peer-to-Peer-electronic-Cash-System.pdf>
- Narayanan A, Bonneau J, Felten E, Miller A, Goldfeder S (2016) Bitcoin and cryptocurrency technologies: a comprehensive introduction. Princeton University Press, Princeton
- Pentland A, Shrier D, Wladawsky-Berger THI (2016) Towards an Internet of trusted data: a new framework for identity and data sharing, Massachusetts Institute of Technology. https://www.nist.gov/sites/default/files/documents/2016/09/16/mit_rfi_response.pdf
- Vukolić M (2015) The quest for scalable Blockchain fabric: proof-of-work vs. BFT replication. Open problems in network security. Springer International Publishing
- Wood G (2014) Ethereum: a secure decentralised generalised transaction ledger. Ethereum project yellow paper, pp 1–32
- Zheng Z, Xie S, Dai H, Chen X, Wang H (2018) An overview of Blockchain technology: architecture, consensus, and future trends. In: Proceedings of the Big-Data congress. IEEE, Seattle

Blockchain-Based Currencies

- [Blockchain](#)

Bluetooth Low Energy (BLE) Security and Privacy

Yue Zhang¹, Jian Weng¹, Rajib Dey³,
and Xinwen Fu^{2,3}

¹College of Information Science and
Technology, Jinan University, Guangzhou,
China

²University of Massachusetts Lowell, Lowell,
MA, USA

³Department of EECS, University of Central
Florida, Orlando, FL, USA

Synonyms

[Provide synonyms to the article title](#)

Definitions

This survey focuses on privacy and security of Bluetooth Low Energy (BLE), covering the attack vectors and corresponding countermeasures.

Bluetooth Low Energy

We are approaching toward a new era of the Internet of things (IoT). Manufacturers are rushing to enrich their products with more IoT functionalities. Bluetooth Low Energy (BLE/BTLE) will act as one of the backbones for IoT wireless communication technologies. Compared with the Bluetooth Classic technology, BLE provides larger coverage area while maintaining a lower power consumption. According to the Bluetooth SIG forecast, over 90% mobile devices will support BLE by 2018 (Group 2014).

Bluetooth Low Energy (BLE) is part of Bluetooth Core Specification and is a wireless technology specifically designed to be used for novel applications in IoT. It was released in 2010 along with Bluetooth 4.0. Since then BLE 4.1, BLE 4.2, and BLE 5 (Wikipedia 2018) have been released alongside its classic counterpart.

BLE is not compatible with Bluetooth Classic. Thus, a device implementing only the low energy feature cannot communicate with a device that only implements Bluetooth Classic (Gomez et al. 2012). Other differences between BLE and Bluetooth Classic are listed in Table 1.

In the rest of this survey, we will discuss two main protocols of BLE, its privacy features, and the technical details that protect the BLE privacy. We will discuss the security issues of BLE, involving security goals and association models for secure communication. We will introduce major attacks on BLE protocols, such as DOS attack, eavesdropping, and the man-in-the-middle (MITM) attack. We will discuss vulnerabilities and design issues that cause these attacks. Finally, we will give suggestions on how to make BLE device more secure.

ATT and GATT Protocol

The ATT protocol is a client-/server-based protocol (Bluetooth 2014). The server maintains a set of attributes which can be accessed by the client once connected. Each of these attributes has a handle, a type with a universally unique identifier (UUID), a value, and a set of permissions. The handle is a 16-bit, non-zero value that is assigned by the server to uniquely index an attribute. The type is defined by UUID. Given the burden of transferring UUID values, the typical 128-bit UUID is converted to the 16-bit UUID in BLE. The UUID indicates the functionality of an attribute. Permissions restrict the accessibility of each attribute for another device, such as *read/write*, and *notify*. Table 2 is an example that demonstrates an attribute.

The GATT protocol is based on the ATT protocol and organizes attributes into *services* (Townsend et al. 2014). A service contains zero or several *characteristics*. Each characteristic contains zero or several *descriptors* as well. Characteristics are containers for user data. These descriptors are attributes that convey other information about a characteristic and its value. Services may include other services as their building blocks. The major service that contains other services is called the primary service, while the auxiliary ones are named secondary services.

Bluetooth Low Energy (BLE) Security and Privacy, Table 1 Differences between BLE and Bluetooth Classic (Wireless World 2018)

Features	Bluetooth Classic	BLE (Bluetooth Low Energy)
Range	<30 m	More than 50 m
Power consumption	30 mA	Less than 15 mA
Speed	700 Kbps	1 Mbps
Frequency channels	79 channels	40 channels (includes 37 data channels and 3 advertising channels)
Modulation	GFSK (modulation index 0.35)	GFSK (modulation index 0.5)
Latency in data transfer	100 ms	3 ms
Message size(bytes)	358 (Max)	8–47
Throughput	0.7–2.1 Mbps	Less than 0.3 Mbps
Slaves	7	Unlimited

Bluetooth Low Energy (BLE) Security and Privacy, Table 2 Attribute for Device Name

Handle	Type (UUID)	Value	Permissions
0 × 0001	Device Name (0 × 2900)	Smart Lock	Read

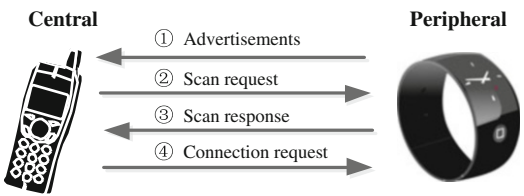
Secondary services must be included in primary services.

Connection Setup

There are two roles in the connection setup procedure. The client is named central device. The central device can be some smart terminals that often run on an operating systems. For example, our smartphone often plays the role of a central device. The server is named peripheral device. A peripheral device is normally a device that has specific functions, such as a smart thermometer, a smart light, or a smart lock. Sometimes, our smartphone can also be a peripheral, as long as it can provide services to some other devices. A peripheral can advertise and notify other devices that it is there. The central device responds to these advertisements with a scan request and then sets up a connection. Note that the advertisements and scan modes vary from each other (Bluetooth 2014).

We elaborate the connection setup procedure in Fig. 1.

1. The peripheral device broadcasts *advertisement* packets indicating it is connectible. These advertisement packets often contain



Bluetooth Low Energy (BLE) Security and Privacy, Fig. 1 Initial states of Bluetooth Low Energy connection

- the basic information about the device, such as the manufacturer.
2. The central device receives these advertisements and then sends a *scan* request to the peripheral to query more information about the peripheral.
 3. The peripheral device responds with a *scan response* packet that may contain more detailed information of itself, such as device name.
 4. After handling the *scan response*, the central then sends the *connect request*, if the response is in line with its expectations. The *connect request* includes the channel information that these two devices will follow during the communication.

Once a connection is established, the central device becomes the **master**, while the peripheral device becomes the **slave**. The **master** may have several **slaves** at the same time, while the **slave** can only have one **master** in 4.0 (Bluetooth 2010). This situation changes in the 4.1 (Bluetooth 2013), which allows one **slave** to have links to more than one **master** at a time.

Privacy of Bluetooth Low Energy

Privacy in BLE is an ability that avoids one device being tracked by other untrusted devices. This type of ability depends on if the device can prevent its address being decoded by other unauthorized devices. If this address gets exposed, other devices may follow the address from the advertising phase. To maintain the privacy, two devices share a secret key known as the Identity Resolving Key (IRK). The IRK is used to generate a random address that only can be resolved by the peer device who is maintaining the IRK. In the rest of this section, we will describe this process in detail.

Bluetooth Device Address

Each Bluetooth device is allocated a uniquely 48-bit Bluetooth device address (BD_ADDR). These addresses are classified as public device addresses and random device addresses. The public device address consists of two 24-bits long values, known as a company ID and a company-assigned ID. The random device address is a randomly generated address. It can be either static random address or private random address. The static random address does not change in one power cycle, while the private random address can change in one power cycle. The private random address may change at every connection. In terms of the private random address, two subtypes of addresses are included, non-resolvable private address and resolvable private address. The resolvable private address (RPA) is the foundation for privacy in BLE.

RPA Generation and Resolution

A RPA is generated from a random number (known as prand) and an Identity Resolving Key (IRK) which is shared during the pairing process. The RPA is used by one device, and only the

corresponding device that maintains the same IRK could resolve it. The format of a RPA is shown in Fig. 2.

Assume that two devices already have the IRK, respectively. To generate a resolvable private address, a device first generates a 24-bit random value (known as prand) and then passes this value with the aforementioned IRK to a hash function. The random address is concatenated by the prand and the output of this hash function, as shown in Eq. (1). Correspondingly, once another device receives this RPA, it splits it to extract the hash value and prand. Then, it performs the hash calculation as well as the first device. The RPA is resolved only when the output matches the extracted hash value.

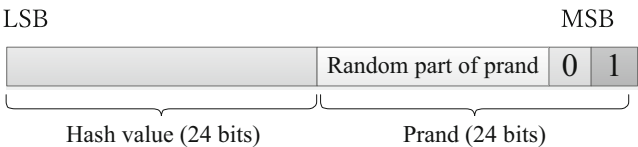
$$Hashvalue = Hash(Prand, IRK)$$
$$RPA = Hashvalue || Prand, \quad (1)$$

Other Policies

Device filtering is a set of policies that avoid untrusted connections. It is also an effective way to reduce power consumption. There are three types of policies: advertising filter policy, scanner filter policy, and connection initiator filter policy. Suggested by their name, these policies affect different states before setting up the connection. When the policies are enabled, a device does not need to respond to every peer device, since most of the devices are filtered out. To do this, a device should maintain a white list locally, which stores the address and its type. For example, the advertising filter policy determines in which way could an advertiser process the scan/connection request. When the specific advertising filter policies are applied, devices may only allow scan/connection requests from whitelist.

The directed connectible advertisement is another way to protect the privacy of the device

Bluetooth Low Energy (BLE) Security and Privacy, Fig. 2 The format of resolvable private address



and avoid unwanted scan or connection requests. This special type of advertisement contains a RPA for the device it intends to connect to. When this type of advertisement is enabled, the device shall only share its information with the ones it wants to connect to.

Security of Bluetooth Low Energy

The application of the BLE technology will continue to grow exponentially in coming years. It also draws attackers attention at the same time. Thus, the security of BLE will have colossal importance.

The security architecture of Bluetooth has evolved over time. Initially, Bluetooth was not as secure as it is today. It could only provide weak protection for the message integrity and can be easily forged. After Version 2.1 + EDR, the Secure Simple Pairing (SSP) was introduced to mitigate the situation. This was a great revolution for Bluetooth Security. Four association models, which are practiced as of now, were also introduced at this version, known as Just Works, Numeric Comparison, Passkey Entry, and Out-Of-Band (OOB).

Along with the release of Version 4.0, the Bluetooth Special Interest Group (Bluetooth SIG) added the security model for Low Energy. However, only three types of association models were supported at this time (Just Works, Passkey Entry, and Out-Of-Band). Although they have the same name as SSP, the security provided is not the same. Some of these association models have been proved to be vulnerable.

In Version 4.1, Bluetooth SIG upgraded SSP by supporting the P-256 elliptic curve and added FIPS-approved algorithms for device authentication, which was known as Secure Connections feature. They also added message integrity by supporting AES-CCM. But they did not change the security features for the Low Energy counterpart.

It is generally acknowledged that the updates in Version 4.2 are crucial for BLE, which greatly enhances its security. At first, it added the Secure Connections feature to Low Energy,

such as upgrading the LE pairing model by supporting the AES-CMAC and P-256 elliptic curve algorithms. Secondly, it included the Numeric Comparison as a new association model in Secure Connections. Lastly, it also applied the Secure Connections features to key generation.

Security Goals

To maintain the security in Bluetooth Low Energy wireless technology, the designers must address two security goals: protection against passive eavesdropping attacks and protection against man-in-the-middle (MITM) attacks. Passive eavesdropping attacks occur when a user is communicating with the other, while an attacker is listening to their conversation stealthily without their consent. The MITM attack is a scenario where the attacker secretly emulates the partner of both communicating devices during a conversation. The attacker here has the ability to listen and change the data of that private communication.

To defend against passive eavesdropping, the traffic between two devices must be encrypted. In the earlier versions of BLE (4.0 and 4.1), the encryption is vulnerable to a brute-force attack. The Temporary Keys (TKs), which are used for further key generation and distribution, are not strong enough. However, this situation was mitigated in the newly released BLE Secure Connection. Although passive eavesdropping can be thwarted by a strong link layer encryption, the link layer encryption may be vulnerable to MITM attacks. Some association models can effectively defend against MITM attacks.

Association Models

Association models are mechanisms that determine how two devices could authenticate each other and securely exchange data with each other. In BLE, association models are used during the pairing process. As of now, there are four association models. They are Just Works, Numeric Comparison, Passkey Entry, and Out-Of-Band (OOB). Which type of the association model would be used depends on the I/O capabilities of both devices. Table 3 shows the mapping of association models according to different I/O capabil-

Bluetooth Low Energy (BLE) Security and Privacy, Table 3 How association model can be used according to different I/O capabilities

Association models \ I/O	Display	Display, Yes/No	Keyboard	No input/output	Keyboard, display
Display	Just works	Just works	Passkey entry	Just works	Just works
Keyboard only	Passkey entry	Passkey entry	Passkey entry	Just works	Passkey entry
No input/output	Just works	Just works	Just works	Just works	Just works
Keyboard, display	Passkey entry	Numeric comparison	Passkey entry	Just works	Numeric comparison

B

ities. The rows and the columns represent the I/O capabilities of two devices, respectively. Display stands for whether the device is equipped with a display, such as a screen. While keyboard stands for whether it has a keyboard. Yes/no here can be a space key, which could be used by the user to confirm or reject a value displayed on a screen.

Now we will discuss these association models in detail. We will focus whether these models could or could not cover the security goals mentioned earlier:

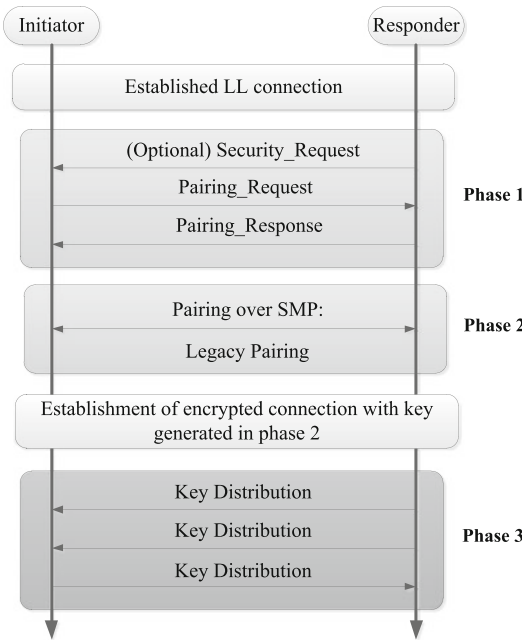
- **Numeric Comparison (only for LE Secure Connections):** Numeric Comparison is applicable when both devices have the capability to display and confirm information. For example, two smartphones want to pair with each other. In this association model, both devices display a six-digit number on its screen. If these two numbers are the same, they are authenticated properly. The numbers are generated from the public key of each device, a nonce, and an 8-bit code by using the AES-CMAC function. During MITM attacks, public keys of both devices will not match. As a result, the numbers displayed on each device will not match. In this way we would avoid MITM attacks effectively.
- **Just Works:** Just Works is designed for scenarios where at least one of the devices has no input/output capability. For example, a mouse and a laptop want to pair with each other. In this association model, they use the same protocol as the Numeric Comparison, but the six-digit number never shows up. Because Just

Works does not provide an interface for a user to know if the number matches, it is vulnerable to MITM attacks.

- **Passkey Entry:** Passkey Entry is designed for scenarios where one device supports entering the six-digit number (Passkey), but does not have the capability to display. For example, a keyboard and a laptop want to pair with each other. In this association model, the final confirm value is generated from the public key of both devices, a nonce and the one part of the Passkey by using AES-CMAC function.
- **Out-of-Band (OOB):** Out-of-Band is designed for scenarios where both devices possess an extra communication channel, such as WLAN or NFC, for example, two phones equipped with NFC. In this association model, the capability to defend against MITM attacks or passive eavesdropping attacks depends on what type of communication channel being used and security of that communication channel.

Pairing and Encryption

Pairing occurs when two BLE devices build a connection, which is a process used for key exchanging. Pairing involves three phases. One of the association models is used for authentication during the pairing process. Depending on the version of Bluetooth Low Energy, there are two types of pairing methods. In 4.1 and 4.0,



Bluetooth Low Energy (BLE) Security and Privacy, Fig. 3 LE legacy pairing

LE legacy pairing is used, while LE Secure Connections are supported in 4.2.

LE Legacy Pairing

We first introduce LE legacy pairing. It contains three steps, as shown in Fig. 3. In the first step, both devices exchange their I/O capabilities and security requirements, which is used to determine the association model. In the second step, the two devices determine a Temporary Key (TK) and use this to generate the Short-Term Key (STK). In the third step, a Long-Term Key (LTK) would be generated for future communications.

Determining Temporary Key (TK) TK is a 128-bit key that is used to generate the Short-Term Key (STK). Recall that three types of association models are supported to share the TK, known as Just Works, Passkey Entry, and OOB. In Just Works, the TK is always set to 0. In Passkey Entry, validated portion of TK is a random number between 000000 and 999999, while other bits can be padded with 0. Only in OOB,

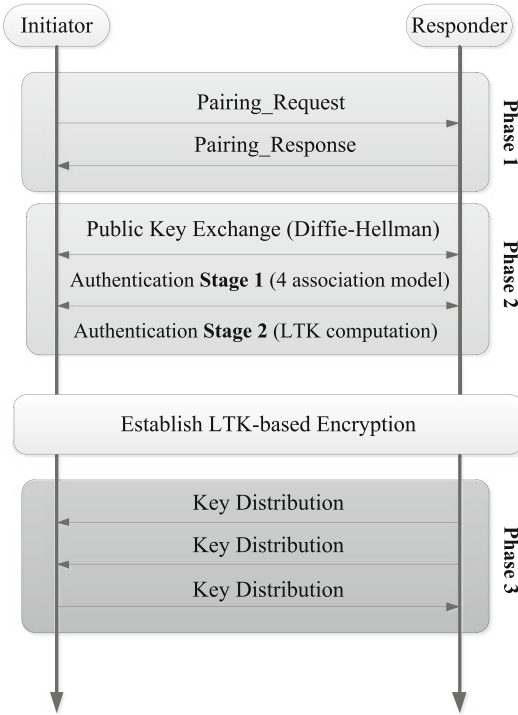
the 128-bit TK can be applied. During the pairing process (Zegeye 2015), few values exchanged in plain text in the air, which can expose the TK to the attacker. These values are confirm value, devices information, pair request and response, Mrand (random value from master), and Srand (random value from slave). By knowing the *confirm value generation function* $c1$ that is defined in 4.0 (as shown in Eq. (2)), attackers could use these values to brute-force the TK in the Passkey Entry model (only 1,000,000 possibilities). Not to mention that in the Just Works model, TK is set to 0.

$$\begin{aligned}
 Mconfirm &= c1(TK, Mrand, \\
 &\quad Pairinginfo, deviceinfo) \\
 Sconfirm &= c1(TK, Srand, \\
 &\quad Pairinginfo, deviceinfo) \quad (2)
 \end{aligned}$$

Determining Short-Term Key (STK) After checking the correctness of the confirm values, devices exchange the generated random numbers (Srand and Mrand). Then the Short-Term Key (STK) is generated using the TK and random numbers (as shown in Eq. (3)).

$$STK = AES128(TK, Srand || Mrand) \quad (3)$$

Determining Long-Term Key (LTK) Devices which support encryption in the Link Layer Connection State in the slave role will generate LTK, the Encrypted Diversifier (EDIV) and 64-bit Random Number (Rand), IRK (Identity Resolving Key), and CSRK (Connection Signature Resolving Key). Among them, the LTK is used to generate session key (SK) for Link Layer Connection, while the IRK is used for random resolvable private address resolution. CSRK supports the data signing operation, which adds another security layer. A signature will be generated with the signing algorithm and a counter by using the CSRK to avoid replay attacks.



Bluetooth Low Energy (BLE) Security and Privacy,
Fig. 4 LE Secure Connections

LE Secure Connections

Compared to the LE legacy pairing, LE Secure Connections offer significantly stronger security. The TK and STK are not used. Instead, a single LTK is used to encrypt the connection, which is exchanged by using the standard (FIPS) compliant Elliptic Curve Diffie-Hellman (ECDH) public key (Sachin Gupta Rohit Kumar 2018). As shown in Fig. 4, we introduce the three steps in LE Secure Connections. In the first step, both devices settle on a certain set of domain parameters. In the second step, the two devices exchange their public keys and compute a Long-Term Key (LTK). In the third step, a set of keys are generated and distributed for future communications.

Public Key Exchange In this step, each device sends its public key to the other device (PK1 and PK2). This public key is generated from the Elliptic Curve Diffie-Hellman (ECDH). It is an effective way to thwart eavesdropping, since the

private key is never exposed in the air during the pairing.

Authentication Stage 1 To perform the authentication, one of the aforementioned association models must be used in the first authentication stage. At this stage, each device will generate a nonce (N1 and N2), which is needed to calculate a commitment (C1 and C2). Taking the Numeric Comparison as an example, the process can be described as Eq. (4). During this stage, N1 and N2 are exchanged, while the generated value D is displayed on each devices. Since there are no extra user inputs in Numeric Comparison, the user input values (ri1 and ri2) are set to 0. The Just Works model follows the same protocol as Numeric Comparison. However, Just Works is still vulnerable to the MITM attack, since there are no user interactions during the whole process. For the OOB or Passkey Entry, the process is similar. For example, in the Passkey Entry model, the ri values are not set to 0 but set to the input values of users. And the last calculation is also more complex.

$$Device1 : C1 = f(PK1, PK2, N1, ri1);$$

$$(ri1 = 0)$$

$$Device2 : C2 = f(PK1, PK2, N2, ri2);$$

$$(ri2 = 0)$$

$$D = g(PK1, PK2, N1, N2) \quad (4)$$

Authentication Stage 2 In this stage, each device collects the exchanged values to generate a Long-Term Key. Both devices will also perform DHKey (Diffie-Hellman key) check process after LTK is generated. Each side uses the DHKey as input to compute new confirmation value by using the previously exchanged information in the DHKey check process. If the check passes through on both sides, the devices start to generate the LTK. LTK is used for the link encryption and sharing keys.

Attacks on Bluetooth Low Energy

It is generally acknowledged that security in LE legacy pairing is vulnerable from many perspectives. Attackers are more likely to choose 4.1 and 4.0 BLE versions as their target as they use the legacy pairing. On the other hand, vulnerabilities can also stem from the implementation or design by any particular device manufacturer.

DoS Attack

A denial-of-service attack (DoS attack) is an attack that renders the network or computing resource unavailable to the legitimate users by disrupting the services. In BLE, the DoS attack is easy to launch and hard to defend.

In Version 4.0, the slaves can only have one master. If the master is connected to the slave, the slave will stop broadcasting the advertisement packets and will become invisible to any other devices. This makes it vulnerable to DoS attacks, since attackers could enhance their advertisement rate to win such a connection race. The situation has been mitigated in Version 4.1 as it supports the multi-master model.

The jamming attack is another DoS attack. It blocks the advertisements of a slave and makes it unconnectable. As an example, Sebastian et al. present a design of selective jammers against BLE advertising, which can block advertisements of a specific slave from being revived by other devices (Brauer et al. 2016). They theoretically prove that the selective jammer is effective and hard to detect.

Another DoS attack is known as denial-of-sleep (DoSL) attack. It can paralyze the device by draining out their battery. In this type of attack, the attackers will consume the power of the device by connecting/disconnecting with them repeatedly (Uher et al. 2016).

Passive Eavesdropping

Passive eavesdropping occurs when two devices communicate with each other and an attacker is listening to their conversation stealthily without their awareness. BLE versions 4.0 and 4.1 are subject to this type of attack, since they use the LE legacy pairing method. Specifically, for the

Just Works and the Passkey Entry of LE legacy pairing, TK is guessable, since the possibility of a TK is very limited. Since BLE uses only 40 frequency channels compared with 79 channels in Bluetooth Classic, BLE traffic is easy to follow. Manufacturers are more devoted to usability than security. Their user guides often do not mention pairing or security even if a more secure pairing method is available. Users may not use any secure pairing at all, and the BLE communication traffic will be in plaintext. Therefore, the attacker can easily deploy attacks such as command replay or injection against their products (Jasek 2016).

In 2011, Michael Ossmann released the first open-source low-cost Bluetooth test tool, named Ubertooth One (Ossmann 2011). In 2013, Mike Ryan built a BLE sniffer over Ubertooth and explored the vulnerability in the LE legacy pairing. A tool called *Crackle* was released at that time to attack Just Works and the legacy Passkey Entry pairing methods (Ryan et al. 2013) of BLE 4.0 and 4.1. In 2014, The Adafruit Bluefruit LE sniffer was introduced Townsend (2014). The Passkey Entry pairing method in BLE 4.2 is secure with a new passkey exchange protocol.

MITM Attack

The MITM attack brings havoc to the BLE protocol. In a MITM attack, the attacker pretends to be a one device (a slave or a master), intercepts the connection from the its peer device, and then relays messages between the two original communicating parties. The Just Works pairing method of any version of BLE suffers from this attack. For example, for LE Secure Connections, Just Works does not check if the six-digit confirm number matches, mostly because there is no I/O such as a display on the devices. Sometimes, even if a victim device has necessary IO capabilities, an attacker could trick them force to use the Just Works. For example, an attacker impersonates as one device and tells the other device that it has no I/O capability. As a result, the only choice is the Just Works pairing method. Therefore, the MITM attack can then be launched (Haataja and Hypponen 2008).

The spoofing attack is a special type of MITM attack or can be regarded as a building block

of the MITM attack. The attacker pretends to be a legitimate device and communicates with the target collecting sensitive information, such as password. That is, in a spoofing attack, the attacker does not relay messages, but pretends to be the slave of interest communicating with the master. Numeric Comparison (only for LE Secure Connections) and Passkey Entry could avoid this type of attack. The OOB pairing can defeat the MITM attack if the OOB channel is secure.

To perform a MITM attack, an attacker needs two role players (i.e., two BLE devices), a DoS attacker and a fake service provider. The DoS attacker blocks the communication between the master and the slave, while the fake service provider forges a fake Bluetooth MAC address that is the same as the victim slave's MAC address, fake advertisements, and fake services. To know all such information, the attacker has to perform sniffing and reverse engineer the applications and even the target hardware sometimes.

Attacks Against BLE Apps

Given the extra cost (such as a display or a Wi-Fi adapter) of using more secure association models and convenience, BLE device designers may prefer to use Just Works. As we know, Just Works is subject to the MITM attack. To address this, the developers of the smart device can implement extra security mechanisms on the application layer, such as password-based or challenge-response authentication.

We now discuss design pitfalls of the application layer security mechanisms.

Mis-Bonding (or Co-located attack in the context of BLE) Under the context of Bluetooth-enabled smartphones, Mis-Bonding refers to the problem that an unauthorized app can access the sensitive user data on the smart device. The data was only designed for the official app initially. Because of vulnerabilities existing in the smartphone's permission mechanism, the data becomes accessible for all apps, even a malicious one. In this case, a malicious app could stealthily gather the data of a device that paired with the smartphone like an official app and then fed the

data to a remote malicious server (Naveed et al. 2014). Because this is a vulnerability in application layer, the link layer encryption is useless in this case.

Password authentication Many BLE-enabled devices use the password authentication. The implementation is often problematic.

- **Clear-text data transformation:** Passwords transmitted in plaintext can be captured through passive eavesdropping.
- **Unlimited login attempts:** This allows the attacker to launch a brute-force attack. However, there is a trade-off if the number of login attempts is limited. Limiting login attempts allows the attacker to perform DoS attacks.
- **Small password space:** Many devices only support six-digit passwords, since a complex and long password may reduce usability. This makes the devices suffer from brute-force attacks. Designers use specific a hash function or an encoding function to make up the shortness, but finally proved to provide few valid mitigations for this. Attacker can also use multipliable devices in parallel to perform an off-line brute-force attack.

Bluetooth Low Energy v.s. Bluetooth Classic

In Dunning (2010), John Paul et al. summarized the basic attacks in Bluetooth Classic. We add new attacks into the table and try to identify the difference between BLE attacks and attacks against Bluetooth Classic. As shown in Table 4, most attacks against Bluetooth Classic are also threats against BLE.

Defense Measures

Based on the discussion in this paper, we present guidelines for BLE security as follows.

- **Prevent advertisement abuse:** Do not put any sensitive information into broadcast packets.

Bluetooth Low Energy (BLE) Security and Privacy, Table 4 Bluetooth Classic attacks vs. these in BLE

Attack classification	attack description	Threat level	Examples	If exists in BLE
Surveillance	Gather user information (not privacy)	Low	Blueprinting (tsb 2017a)	Y
Range extension	Extend the attack range	Low	BlueSniping (tsb 2017c)	Y
Obfuscation	Hide attacker's identity	Low	bdaddr (Whigu 2018)	Y
Fuzzer	Submit nonstandard input to achieve malicious result	Medium	BlueSmack (tsb 2017b)	Y (Sullivan 2015)
Sniffing	Passive eavesdropping	Medium	FTS4BT (fte 2018)	Y
DoS	Block the services	Medium	Signal jamming (Köppel 2013)	Y
Brute-force address	Brute-force the address of devices to trace connection	Medium	Address exhaustion (altCross et al. (2007))	Y
Brute force in pairing	Brute force in the temporary key or PIN during the pairing	High	btpincrack (Project 2017)	Y
Mis-Bonding	Unauthorized app can access the sensitive data from smartphone	High	DMB (Naveed et al. 2014)	Y
Relay attack	With the help of relay device, implement unwanted unlock	High	Home lock attack (Ho et al. 2016)	Y
MITM	Attacker mask itself as parter to achieve malicious result	High	bthidproxy (Mishra 2012)	Y
Malware	Malwares that propagated through Bluetooth	High	CommWarrior (Wei et al. 2008)	Y
Unauthorized access	Control the smartphone and steal sensitive data	High	Car Whisperer (tsb 2017d)	Y

- **Enforce the user to pair:** This is an effective way to thwart the passive eavesdropping.
- **Use version 4.2 and above:** BLE 4.1 and 4.0 suffer from the TK brute-force attack.
- **Do not use the Just Works if possible:** Just Works is vulnerable to MITM attacks, even in BLE 4.2.
- **Implement encryption in application layer** Designers should not invent their own cryptographic algorithms and protocols if there are well-known ones that can do the job. Encryption should be adopted to prevent passive monitoring data leaks.
- **Strengthen the password authentication:** Increasing the password space and limiting the login attempts are the most straightforward but effective ways to strengthen password authentication, and the communication should be encrypted.
- **Protect source code carefully:** It can be seen that a lot of attacks can be launched partially because attackers are able to reverse engineer the BLE apps (Cyr et al. [2014](#); Rieck [2016](#); Allan and Mistry. [2014](#)). Therefore, obfuscation, JNI, code encryption, and other methods should be applied to protect source code.
- **Implement OTA (over-the-air) capabilities into each device:** Enough space should be allocated in the device memory so that code can be updated to patch holes.

Conclusion

This survey focuses on privacy and security of Bluetooth Low Energy (BLE), covering the attack vectors and corresponding countermeasures. Given the vulnerabilities in BLE 4.0 and

4.1, we highly recommend that BLE 4.2 and 5 should be adopted whenever possible. The Just Works pairing strategy in all BLE versions are subject to MITM attacks. It should not be used. In case that Just Works has to be used because of cost and usability issues, application level encryption and authentication is recommended.

References

- Allan A, Mistry S (2014) Reverse engineering the estimate. <http://makezine.com/2014/01/03/reverse-engineering-the-estimate/>. Accessed 26 June 2018
- Bluetooth S (2010) Bluetooth core specification version 4.0. Specification of the Bluetooth System
- Bluetooth S (2013) Bluetooth core specification version 4.0. Specification of the Bluetooth System
- Bluetooth S (2014) Bluetooth core specification version 4.2. Specification of the Bluetooth System
- Brauer S, Zubow A, Zehl S, Roshandel M, Mashhadi-Sohi S (2016) On practical selective jamming of Bluetooth low energy advertising. In: 2016 IEEE conference on standards for communications and networking (CSCN). IEEE, pp 1–6
- Cross D, Hoeckle J, Lavine M, Rubin J, Snow K (2007) Detecting non-discoverable bluetooth devices. In: International conference on critical infrastructure protection. Springer, pp 281–293
- Cyr B, Horn W, Miao D, Specter M (2014) Security analysis of wearable fitness devices (fitbit). Massachusetts Institute of Technology, p 1
- Dunning J (2010) Taming the blue beast: a survey of Bluetooth based threats. *IEEE Secur Priv* 8(2):20–27
- fte (2018) Fts4bt™ bluetooth® protocol analyzer and packet sniffer. <http://www.fte.com/products/FTS4BT.aspx>
- Gomez C, Oller J, Paradells J (2012) Overview and evaluation of Bluetooth low energy: an emerging low-power wireless technology. *Sensors* 12(9):11734–11753
- Group BSI (2014) Mobile telephony market. <https://www.bluetooth.com/news-events/news>. Accessed 26 June 2018
- Haataja KM, Hypponen K (2008) Man-in-the-middle attacks on Bluetooth: a comparative analysis, a novel attack, and countermeasures. In: 3rd international symposium on communications, control and signal processing, 2008. ISCCSP 2008. IEEE, pp 1096–1102
- Ho G, Leung D, Mishra P, Hosseini A, Song D, Wagner D (2016) Smart locks: lessons for securing commodity internet of things devices. In: Proceedings of the 11th ACM on Asia conference on computer and communications security. ACM, pp 461–472
- Hruska J (2017) Bluetooth low-energy technology isn't just another Bluetooth revision its a whole new technology. <http://suo.im/4KhIEQ>. Accessed 26 June 2018
- Jasek S (2016) Gattacking Bluetooth smart devices. In: Black Hat USA conference
- Köppel S (2013) Bluetooth jamming. ETH Zürich, Zürich
- Mishra N (2012) Defence against man in middle attack in secure simple pairing. *J Glob Res Comput Sci* 3(5): 78–82
- Naveed M, Zhou Xy, Demetriou S, Wang X, Gunter CA (2014) Inside job: understanding and mitigating the threat of external device mis-binding on android. In: NDSS
- Ossmann M (2011) Project Ubertooth: building a better Bluetooth adapter. <http://ossmann.blogspot.com/2011/02/project-ubertooth-building-better.html>. Accessed 26 June 2018
- Project TO (2017) btpincrack. <http://openciphers.sourceforge.net/oc/btpincrack.php>. Accessed 26 June 2018
- Rieck J (2016) Attacks on fitness trackers revisited: a case-study of unfit firmware security. arXiv preprint arXiv:160403313
- Ryan M et al (2013) Bluetooth: with low energy comes low security. *WOOT* 13:1–7
- Sachin Gupta Rohit Kumar (2018) Ble v4.2: creating faster, more secure, power-efficient designs—part 4. <http://www.electronicdesign.com/communications/ble-v42-creating-faster-more-secure-power-efficient-designs-part-3>
- Sullivan H (2015) Security vulnerabilities of Bluetooth low energy technology (BLE). Tufts University
- Townsend K (2014) How the Bluetooth sniffing on android works. Sebastopol, California. <https://learn.adafruit.com/introducing-the-adafruit-bluefruit-le-sniffer/introduction>
- Townsend K, Cufi C, Davidson R, Davidson A (2014) Getting started with Bluetooth low energy. O'Reilly Media, Inc.
- trifinite (2017) Bluesnarf. https://trifinite.org/trifinite_stuff_bluesnarf.html. Accessed 26 June 2018
- tsb (2017a) What is blueprinting in bluetooth security? <https://www.thesecuritybuddy.com/?s=Blueprinting>. Accessed 26 June 2018
- tsb (2017b) What is bluesmack attack? <https://www.thesecuritybuddy.com/?s=BlueSmack>. Accessed 26 June 2018
- tsb (2017c) What is bluesniping? <https://www.thesecuritybuddy.com/bluetooth-security/what-is-bluesniping/>. Accessed 26 June 2018
- tsb (2017d) What is car whisperer? <https://www.thesecuritybuddy.com/bluetooth-security/what-is-car-whisperer/>. Accessed 26 June 2018
- Uher J, Mennecke RG, Farroha BS (2016) Denial of sleep attacks in Bluetooth low energy wireless sensor networks. In: Military communications conference, MILCOM 2016-2016 IEEE. IEEE, pp 1231–1236
- Wei X, Li Z, Chen Z, Yuan Z (2008) Commwarrior worm propagation model for smart phone networks. *J China Univ Posts Telecommun* 15(2): 60–66
- Whigu (2018) Change your Bluetooth device MAC address. <http://blog.petrilopia.net/linux/change-your->

[bluetooth-device-mac-address/](#). Accessed 26 June 2018

Wikipedia (2018) Bluetooth low energy. https://en.wikipedia.org/wiki/Bluetooth_Low_Energy. Accessed 26 June 2018

Wireless World R (2018) Bluetooth vs BLE-difference between Bluetooth and BLE(bluetooth low energy). <http://www.rfwireless-world.com/Terminology/Bluetooth-vs-BLE.html>. Accessed 26 June 2018

Zegeye WK (2015) Exploiting Bluetooth low energy pairing vulnerability in telemedicine. In: International telemetering conference proceedings, international foundation for telemetering

Botnet

► New Variants of Mirai and Analysis

Brain-Machine Interfaces

Josep Miquel Jornet¹, Michal K. Stachowiak², and Sasitharan Balasubramaniam³

¹Ultra-broadband Nano Communication and Networking Laboratory, University at Buffalo, The State University of New York, Buffalo, NY, USA

²Department of Pathology and Anatomical Sciences, University at Buffalo, The State University of New York, Buffalo, NY, USA

³Telecommunication Software and Systems Group, Waterford Institute of Technology, Waterford, Ireland

Definition

Brain-machine interfaces (BMIs) refer to communication systems between the brain and an external device. Desired properties of BMIs include bidirectionality, high spatial and temporal resolution, low invasiveness, accuracy, and robustness. In this paper, the different types of BMIs, the state of the art, and the

future directions are discussed, in addition to highlighting their key applications.

Historical Background

For many decades, the interaction between humans and machines has been restricted to the exchange of visual, auditory, and tactile information. A conceptual analysis of the existing human-machine interfaces (HMI) reveals that the amount of useful information that can be exchanged between humans and machines is not limited by the capabilities of the human brain or those of the machine processor, but by the interfaces between them. Simply stated, from an engineering perspective, the human being can be modeled as a *macro-system* with a processing powerhouse, i.e., the brain, and a collection of peripherals, i.e., the sense organs. All the peripherals have their own latency limitations, which mainly arise from the fact that they convert a collection of *nano-/micro-events*, i.e., neuronal activity in the form of action potential signals, into a *macro-sized effect*, e.g., moving the fingers to touch a display or type some characters in order to control an external machine or, reciprocally, convert a macro-size effects, e.g., an auditive or visual command, into a collection of nano-/micro-events in the brain.

In order to overcome the limitations of the peripheral system, direct communication with the brain is needed. This form of interaction is known as a brain-machine interface (BMI). Over the years, numerous applications have resulted from these two-way interactions, where compensations are made between the computational capabilities of both systems. Such compensations can be in the form of either the computing system analyzing the brain signals and adapting the computing environment or the computing system providing added computing power to compensate for shortcomings of the brain. In the first case, brain signals are used to understand the user's context which is then used to control machines, such as controlling the vehicle. In the second case, this could come in the form of computing systems

controlling neuroprosthetic devices for disabled patients.

Electrical Brain-Machine Interfaces

The most common to date BMIs rely on the collection and excitation of electrical signals from the brain. Electroencephalogram (EEG) signals have been successfully utilized to directly control machines without the need of the sense organs (Millan et al. 2004). EEG signals can be collected in a noninvasive way, i.e., from outside the brain, and support high-temporal resolution, i.e., down to the sub-millisecond scale, but have limited spatial resolution, i.e., cannot be utilized to read the action potential signal from a single neuron at a time, and are vulnerable to electrical artifact sources. Besides EEG-based BMIs, there are other more invasive mechanisms that could be utilized to enable more robust electrical BMIs, such as intracranial EEG (Leuthardt et al. 2006), which is also known as electrocorticogram, and microelectrode arrays, which are placed directly on the exposed surface of the brain (Hochberg et al. 2006). However, besides their invasiveness, they suffer from several limitations, such as complex application or unsuitability for long-term use.

Optical Brain-Machine Interfaces

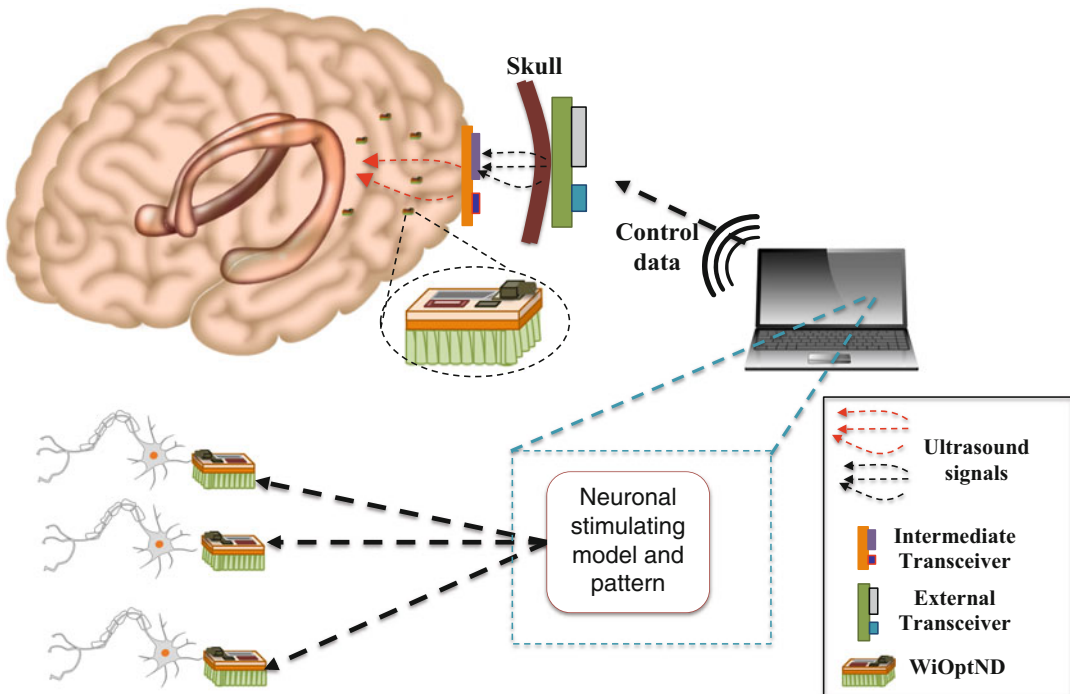
In parallel to the development of the aforementioned approaches, the field of optogenetics, i.e., the use of light to interact with genetically modified neurons in the brain (Zemelman et al. 2002; Deisseroth 2011; Zhang et al. 2007), has experienced a major revolution in the last decade. Optical neural stimulation is considered to be more beneficial than electrical neural stimulation, because it permits activation or inhibition of specific types of neurons with sub-millisecond temporal precision and eliminates electrical artifacts. Current approaches to optogenetic neural interfaces include the use of optical fibers coupled to lasers or light-emitting diodes (LEDs) (Zorzos

et al. 2010) and micro-LED arrays (McGovern et al. 2010). Moreover, optogenetics enables bidirectional interfaces, as light can be utilized both to control and to measure neuronal activity (Kwon et al. 2014). However, the size of existing optical devices makes them invasive, difficult to contact to individual neurons, and, ultimately, not suitable for chronic BMIs (Marblestone et al. 2013).

Wireless Brain-Machine Interfaces

In order to overcome the limitations of traditional electrical and optical BMIs, wireless BMIs are being developed. In Seo et al. (2013), the concept of neural dust was introduced for the first time. In the envision architecture, miniature electronic devices or dust motes are implanted in the cortex. These devices, which integrate piezoelectric energy harvesting systems powered by ultrasounds, record the neural activity from the cortex and transmit the information to a subdural transceiver mounted under the skull. This device is in charge of controlling the neural dust and to communicate with the external head-mounted transceiver, where the data is collected. Despite the advantages of this wireless architecture, the fact that it relies on the principles of electrical BMIs limits its applications.

Recently, in Wirdatmadja et al. (2017), the first wireless BMI based on wireless optogenetic nanonetworking devices (WiOptNDs) was proposed (Fig. 1). WiOptND enables accurate, robust, high-throughput, and minimally invasive BMIs by leveraging the state of the art in nanophotonics, nanoelectronics, and wireless communication. The fundamental idea is to replace existing micro-LED arrays and micro-photodetector arrays used in optical BMIs by a network of coordinated nano-devices, which are able both to excite individual neurons and to measure their activity. In this application, a network of collaborative WiOptNDs is utilized both to excite multiple neurons according to incoming commands and to collect, process, and transmit accurate neuronal activity in real time.



Brain-Machine Interfaces, Fig. 1 The WiOptND architecture consists of (i) a network of coordinated nano-devices able to optogenetically excite and measure the response of neurons; (ii) an intermediate transceiver in charge of both controlling the nano-devices in order

to generate different neuron excitation patterns as well as acoustically powering them; and (iii) an external transceiver in charge of acoustically powering the intermediate transceiver and interfacing it with the actual BMI user

Each nano-device is equipped with an optical nano-transceiver (Feng et al. 2014) and nano-antenna (Nafari and Jornet 2017), which is able to both emit and detect optical radiation at a pre-established frequency or wavelength. As in Seo et al. (2013), WiOptNDs are acoustically powered and remotely controlled through the subdural transceiver (Fig. 1).

Many benefits in this approach exist. First, the very small size of optical nano-antennas (Dorf-muller et al. 2010; Nafari and Jornet 2017), below 1 micrometer in the largest dimension, enables the possibility to measure the neuronal activity in a single neuron, with very high accuracy. Moreover, the total size of each individual nano-device, up to a few cubic micrometers at most (Akyildiz and Jornet 2010), minimizes the invasiveness of this approach when compared to

existing optogenetic approaches, which require bulky lasers or optical fibers. Moreover, by operating at optical frequencies, much higher temporal resolution than traditional electrical BMIs can be achieved. For example, while the main features of action potential signals are in the millisecond scale, the possibility to measure those signals with much higher temporal resolution, such as a few microseconds or even less, may unveil new high-frequency time transients in the action potential signal propagation, which could shine new light into the exploration of neuronal pathways. This also enables potentially much faster BMIs. For the time being, however, electrical and optogenetic BMIs are at an early stage, in which some of the system components have been developed and tested, but a fully functional BMI has not been realized.

Future Directions

To enable practical long-term implantable wireless BMIs, there are several bottlenecks that need to be overcome.

- From the *hardware perspective*, the major challenges are in terms of the miniaturization of the neural dust motes. Nano-lasers, nano-antennas, and nano-photodetectors are needed to excite and monitor neural activity through optogenetics. For the time being, the smallest laser, experimentally demonstrated to date (Feng et al. 2014), is a micro-ring laser with 10 μm in diameter. Given the average size of a neuron cell body, tens of micrometers, current nano-lasers should be able to achieve single neuron resolution, provided that they are near it (otherwise, light spreading would result into the illumination of multiple neurons). Optical nano-antennas can be utilized to then overcome this problem. Besides the optics, piezoelectric energy harvesting nano-systems (Wang 2008; Wang et al. 2017) and minimal computational and data storage capabilities are needed.
- From the *communications perspective*, effective communication protocols to operate the WiOptNDs are needed. Addressing of individual nano-devices, precise triggering of the optical stimulation, and accurate collection of information are crucial tasks to be performed reliably and with energy efficiency in mind. When developing such protocols, the physics of the intra-body channel, which affects the propagation of optical signals for optogenetic excitation and measurement (Wirdatmadja et al. 2018) as well as of the acoustic signals required to power the nano-devices (Donohoe et al. 2016), need to be taken into account. As important as their impact on the signal power, their impact on time distortion and synchronization needs to be studied (Noel et al. 2018).
- In all cases, *human factors* need to be taken into account. For the time being, the advanced BMIs (i.e., anything beyond EEG signal col-

lection from over the skull) have been primarily tested in vitro cell cultures or animals. One of the promising approaches relies on the use of cerebral organoids, i.e., artificially grown, in vitro, miniature organs resembling the brain. These organoids reproduce the exact behavior of the brain and are the basis of recent breakthroughs in neuroscience (Stachowiak et al. 2017). Beyond in vitro testing, the implantation of the devices needs to be optimized. Beyond surgery, the combination of nasal injection and self-assembly techniques for nano-devices is being considered.

Key Applications

- BMIs can significantly improve the quality of life of people with disabilities, by providing them a transformative way to interact with the environment and restoring functional abilities and even cognition. For example, the direct control of machines from the brain can help to overcome the limitations in the “interfaces” between them, namely, the sensor organs or locomotion apparatus.
- BMIs can help to broaden the understanding of the developmental- and aging-related diseases, such as schizophrenia or Alzheimer, whose origin lies at communication problems between consecutive neurons, and, ultimately, enable transformative treatments.
- BMIs can help to control specific types of neurological disorders, which to date have been a major challenge. A good example is epilepsy. The development toward using miniaturized neural dust motes that can be implanted deep into the brain, coupled with optogenetics, can provide a mechanism of controlling neurons at a single cell level. This means that by understanding the source of the epilepsy, controls can be developed to suppress the seizure signaling.
- BMIs can help to augment people’s brain power going into the future. New concepts such as *Targeted Neuroplasticity Training* can emerge, whereby the computing power capa-

bilities are augmented with the brain's training process. This could enhance the brain with new skills and also provide people the abilities to acquire new skills where they previously lacked.

Cross-References

- [Nanonetworks](#)
- [Nanoscale Terahertz Communications](#)
- [Terahertz-band nano-communications](#)

Acknowledgments This work was supported by the U.S. National Science Foundation (NSF) under Grant No. CBET-1706050.

References

- Akyildiz IF, Jornet JM (2010) Electromagnetic wireless nanosensor networks. *Nano Commun Netw* (Elsevier) 1(1):3–19
- Deisseroth K (2011) Optogenetics. *Nat Methods* 8(1):26–29
- Donohoe M, Balasubramaniam S, Jennings B, Jornet JM (2016) Powering in-body nanosensors with ultrasound. *IEEE Trans Nanotechnol* 15(2):151–154
- Dorfmueller J, Vogelgesang R, Khunsin W, Rockstuhl C, Etrich C, Kern K (2010) Plasmonic nanowire antennas: experiment, simulation, and theory. *Nano Lett* 10(9):3596–3603
- Feng L, Wong ZJ, Ma RM, Wang Y, Zhang X (2014) Single-mode laser by parity-time symmetry breaking. *Science* 346(6212):972–975
- Hochberg LR, Serruya MD, Friebs GM, Mukand JA, Saleh M, Caplan AH, Branner A, Chen D, Penn RD, Donoghue JP (2006) Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442(7099):164–171
- Kwon KY, Lee HM, Ghovanloo M, Weber A, Li W (2014) A wireless slanted optrode array with integrated micro LEDs for optogenetics. In: 2014 IEEE 27th International Conference on Micro Electro Mechanical Systems (MEMS). IEEE, pp 813–816
- Leuthardt EC, Miller KJ, Schalk G, Rao RP, Ojemann JG (2006) Electrocorticography-based brain computer interface—the Seattle experience. *IEEE Trans Neural Syst Rehabil Eng* 14(2):194–198
- Marblestone AH, Zamft BM, Maguire YG, Shapiro MG, Cybulski TR, Glaser JJ, Amodei D, Stranges PB, Kalhor R, Dalrymple DA, Seo D, Alon E, Maharbiz MM, Carmena JM, Rabaey JM, Boyden ES, Church GM, Kording KP (2013) Physical principles for scalable neural recording. *Front Comput Neurosci* 7(137). <https://doi.org/10.3389/fncom.2013.00137>
- McGovern B, Berlinguer Palmini R, Grossman N, Drakakis E, Poher V, Neil M, Degenaar P (2010) A new individually addressable micro-led array for photogenetic neural stimulation. *IEEE Trans Biomed Circuits Syst* 4(6):469–476
- Millan J, Renkens F, Mourino J, Gerstner W (2004) Noninvasive brain-actuated control of a mobile robot by human EEG. *IEEE Trans Biomed Eng* 51(6):1026–1033
- Nafari M, Jornet JM (2017) Modeling and performance analysis of metallic plasmonic nano-antennas for wireless optical communication in nanonetworks. *IEEE Access* 5:6389–6398
- Noel A, Makrakis D, Eckford AW (2018) Distortion distribution of neural spike train sequence matching with optogenetics. *IEEE Trans Biomed Eng.* <https://doi.org/10.1109/TBME.2018.2819200>
- Seo D, Carmena JM, Rabaey JM, Alon E, Maharbiz MM (2013) Neural dust: an ultrasonic, low power solution for chronic brain-machine interfaces. *arXiv preprint arXiv:13072196*
- Stachowiak EK, Benson CA, Narla ST, Dimitri A, Bayona-Chuye LE, Dhiman S, Harikrishnan K, Elahi S, Freedman D, Brennand KJ, Stachowiak MK (2017) Cerebral organoids reveal early cortical maldevelopment in schizophrenia? Computational anatomy and genomics, role of fgfr1. *Transcult Psychiatry* 7(11):6
- Wang ZL (2008) Towards self-powered nanosystems: from nanogenerators to nanopiezotronics. *Adv Funct Mater* 18(22):3553–3567
- Wang ZL, Jiang T, Xu L (2017) Toward the blue energy dream by triboelectric nanogenerator networks. *Nano Energy* 39:9–23
- Wirdatmadja SA, Barros MT, Koucheryavy Y, Jornet JM, Balasubramaniam S (2017) Wireless optogenetic nanonetworks for brain stimulation: device model and charging protocols. *IEEE Trans NanoBiosci* 16(8):859–872.
- Wirdatmadja S, Johari P, Balasubramaniam S, Bae Y, Stachowiak MK, Jornet JM (2018) Light propagation analysis in nervous tissue for wireless optogenetic nanonetworks. In: Optogenetics and optical manipulation 2018, vol 10482. International Society for Optics and Photonics, p 104820R. <https://doi.org/10.1117/12.2288786>
- Zemelman BV, Lee GA, Ng M, Miesenböck G (2002) Selective photostimulation of genetically charged neurons. *Neuron* 33(1):15–22
- Zhang F, Aravanis AM, Adamantidis A, de Lecea L, Deisseroth K (2007) Circuit-breakers: optical technologies for probing neural signals and systems. *Nat Rev Neurosci* 8(8):577–581
- Zorzos AN, Boyden ES, Fonstad CG (2010) Multi-waveguide implantable probe for light delivery to sets of distributed brain targets. *Opt Lett* 35(24):4133–4135

Brownian Motion

Mohammad Upal Mahfuz

Richard J. Resch School of Engineering,
University of Wisconsin-Green Bay, Green Bay,
WI, USA

Synonyms

Diffusion; Molecular propagation; Random motion; Random walk; Thermal excitation of molecules

Definition

Brownian motion is the random motion of particles, e.g., molecules, suspended in the fluid medium, e.g., liquid and gas, that results from a large number of collisions those particles experience with the fast-moving particles of the fluid medium.

Acronyms

BM	Brownian motion
MC	Molecular communication
MRBP	Molecule-receptor binding process
RN	Receiving nanomachine
TN	Transmitting nanomachine
VRV	Virtual reception volume

Historical Background

Brownian motion (BM) is an important phenomenon that is the basis of diffusion-based propagation of molecules in molecular communication (MC) and, therefore, is the fundamental principle behind diffusion-based MC in the field of nanoscale communication networks, also known as nanonetworks. In the field of natural and applied sciences, BM is also popularly known as random walk motion of particles. The history of BM is quite old.

Random walk motion of particles was first observed by Scottish botanist Robert Brown (1773–1858) who found with a microscope that pollen particles jiggled at a very high speed while they were suspended in water medium (Brown 1828) and hence the name Brownian motion. However, it was Albert Einstein who explained BM with the theory that it is due to an extremely large number of collisions that particles under investigation experience by interacting with the particles of the fluid medium (Einstein 1905). In the field of nanoscale communication networks (Akyildiz et al. 2008; Bush 2010), the concept of BM received a huge amount of interest among researchers (Ahmadzadeh et al. 2017, 2018; Eckford 2007; Gohari et al. 2016; Haselmayr et al. 2017; Jamali et al. 2018; Kadloor et al. 2012; Koo et al. 2016; Nakano et al. 2012, 2013, 2014; Pierobon and Akyildiz 2010, 2011) who referred to the theory of BM as the fundamental principle of the diffusion process (Berg 1993) to investigate the performance of MC systems among nanomachines. Thus, MC based on BM was considered as an option for communications among nanomachines forming nanonetworks (Akyildiz et al. 2008). Some of the works in this field later contributed to the knowledge base while developing the standards for nanoscale communication networks (IEEE 2016).

Foundations

Physical Process

Random walk motion is the fundamental principle of BM. When molecules are suspended in a fluid medium, due to thermal excitation, the molecules collide with one another and thus they propagate in three dimensions (Berg 1993). Such movements of molecules could be either in a truly random direction when there is no external force present in the system or towards a given direction determined by the presence of an external force in the system (Berg 1993). The collisions between molecules occur on a very tiny timescale on the scale of picoseconds (Höfling and Franosch 2013), and thus billions of such collisions can take place within a small

duration of time. When all the factors that can impact the movement of molecules are considered absent, these collisions are truly random and, therefore, molecules undergo diffusion process (Berg 1993; Berg and Purcell 1977; Crank 1975; Einstein 1905) and thus move in a random direction in the fluid (propagation) medium. Random walk of molecules can be characterized by the two major types, namely, continuous-time and discrete-time random walks, distinguished by the step size of the particle between two states and the time the particle remains in one state before it moves to the next state (Metzler and Klafter 2000).

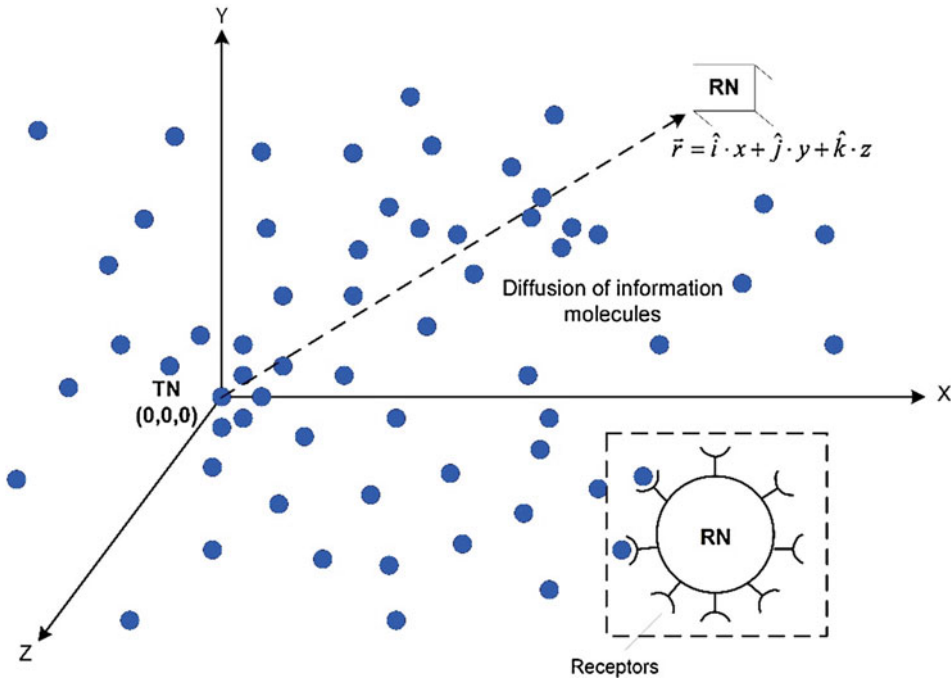
Normal diffusion process is based on uncorrelated random walk model of the movement of the molecule in the fluid medium where the direction of the new movement of the molecule is independent of the direction of its previous movements and can be equally likely in any of the directions in the three-dimensional space. On contrary, anomalous diffusion is based on correlated random walk model of the molecule in the fluid medium where the next movement of the molecule is likely to continue in the current direction (Mahfuz et al. 2016). In the field of MC and nanonetworks, while most researchers have considered normal diffusion process as a result of random walk motion of molecules between a transmitting nanomachine (TN) and a receiving nanomachine (RN), a few researchers have considered anomalous diffusion of molecules and presented investigative results in the aspects of MC systems (Cao et al. 2015; Mahfuz et al. 2016; Mai et al. 2017). While normal diffusion considers random BM of molecules between TN and RN, anomalous diffusion considers the abnormalities involved in the normal diffusion process, for instance, when information molecules are confined within a small space, e.g., intra-cellular space, or when information molecules (particles) interact with one another, e.g., charged particles (Mahfuz et al. 2016). Current literature has also investigated into the role of anomalous diffusion on molecular propagation and availability of molecules at the RN as well as the performance of MC receivers (Cao et al. 2015; Mahfuz et al. 2016; Mai et al. 2017).

Molecular Propagation and Communication

BM is the fundamental principle that is the basis of molecular propagation between a pair of nanomachines in diffusion-based MC systems. As shown in Fig. 1, in the aspects of MC, when a TN, located at the origin (0,0,0) of the axes, emits information-carrying molecules in the fluid medium, these information-carrying molecules experience BM and, after a huge number of collisions with the molecules of the fluid medium, finally reach an RN located at (x,y,z) . As shown in the inset in Fig. 1, the RN is located in the center of a small virtual reception volume (VRV) (Atakan and Akan 2010), which is a small unit volume surrounding the RN where the RN can sense information-carrying molecules, shown as blue circles, that become available to RN and come into contact with the receptors on its surface. Although information molecules can bind with the surface receptors with some probability as governed by the molecule-receptor binding process (MRBP), when a sufficiently large number of surface receptors are assumed, this ensures that the available information molecules at the RN come into contact with the RN and be received by the RN. Thus, the RN receives the information molecules and detects the message by processing the received information-carrying molecules. Here, the RN is assumed to be a transparent receiver that does not affect the BM of the information molecules in the fluid medium. RN is also assumed to perfectly count the number of information molecules that become available within the VRV at a given time.

In the three-dimensional propagation medium denoted by x , y , and z axes, these information-carrying molecules propagate independently in three dimensions and reach the RN located at a distance r from the TN, where $r^2 = x^2 + y^2 + z^2$. The diffusion of information molecules in three dimensions and time (t), in second (s), can be expressed as Eq. (1) below (Berg 1993)

$$\frac{\partial U}{\partial t} = D \nabla^2 U \quad (1)$$



Brownian Motion, Fig. 1 Diffusion of information-carrying molecules in three dimensions in the unbounded propagation medium (Mahfuz et al. 2014)

where U denotes the concentration of information-carrying molecules, i.e., the number of information-carrying molecules per unit volume of reception space, D (in $\mu\text{m}^2/\text{s}$) denotes the diffusion coefficient of information molecules propagating in the given fluid medium, and $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. The value of D is given by $D = \frac{k_B T}{6\pi\eta R_H}$ where $k_B = 1.38 \times 10^{-23}$ J/K is the Boltzman constant, T is the temperature (in K), η is the viscosity of the fluid medium, and R_H is the hydraulic radius of the molecule (Nakano et al. 2013). Diffusion coefficient values of commonly used information molecules can be found in Nakano et al. (2013) and Freitas (1999). Often, it is necessary to measure the values of diffusion coefficient experimentally for which there have been several methods to follow. For instance, several static and dynamic methods to obtain diffusion coefficient values experimentally have been explored in Culbertson et al. (2002). A simple and inexpensive experimental method to measure diffusion coefficient when temperature and pressure vary has been shown in Delgado et al. (2005).

When the TN located at (0,0,0) is assumed to transmit Q_0 molecules, all at a time in an impulsive manner at time $t = 0$, the average concentration of information-carrying molecules, in number of information-carrying molecules per unit volume of the reception space, available at the RN, located at (x, y, z) in an unbounded three-dimensional propagation medium, can be found from the solutions of the diffusion equation in Eq. (1) as follows (Berg 1993):

$$U(r, t) = \frac{Q_0}{(4\pi Dt)^{3/2}} \exp \frac{-r^2}{4Dt} \quad (2)$$

For the transparent receiver mentioned earlier, Eq. (2) shows the average number of information-carrying molecules per unit volume, i.e., concentration of information-carrying molecules, for a given number of transmitted molecules, Q_0 . The concentration of information-carrying molecules $U(r, t)$ depends on distance (r) to the RN and the time elapsed after the release of the molecules from the TN. Apart from the transparent receiver

mentioned in this entry, for other types of MC receivers, interested readers are referred to the following entry of this book: “► [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)” by Ahmadzadeh, A., Jamali, V., Wicke, W., and Schober, R.

There are several simulation frameworks currently available that help researchers to simulate BM in the study of MC-based nanonetworks. For instance, N3Sim uses diffusion-based MC to provide a simulation framework for nanonetworks with transmitters, receivers, and harvesters (Llatser et al. 2014). N3Sim not only uses BM dynamics to model the movements of molecules but also includes inertia of and interaction among the molecules in the system. Another simulation framework named AcCoRD (Actor-based Communication via Reaction Diffusion), which is an open-source project written in C programming language, is designed and used for MC analysis and reaction diffusion solutions (Noel et al. 2017). NanoNS, built on the widely used network simulator NS-2, is another simulation framework that simulates the diffusion-based MC in nanonetworks (Gul et al. 2010). To simulate the behavior of diffusion-based MC with drift effect inside blood vessels, a software simulation platform named BiNS2 has also been introduced (Felicetti et al. 2013).

Key Applications

The applications of BM include all scenarios where particle movement is used. One of the notable applications of BM would be in the diffusion-based MC and nanonetworks fields, for instance, in drug delivery and disease treatment as well as environmental monitoring, protection, and control (Akyildiz et al. 2008; Nakano et al. 2013).

Cross-References

- [Molecular Communication for Wireless Body Area Networks](#)
- [Nanonetworks](#)

- [Reaction-Diffusion Channels](#)
- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

References

- Ahmadzadeh A, Jamali V, Noel A, Schober R (2017) Diffusive mobile molecular communications over time-variant channels. *IEEE Commun Lett* 21(6): 1265–1268
- Ahmadzadeh A, Jamali V, Schober R (2018) Stochastic channel modeling for diffusive mobile molecular communication systems. *IEEE Trans Commun* 66(12):6205–6220
- Akyildiz IF, Brunetti F, Blazquez C (2008) Nanonetworks: a new communication paradigm. *Comput Netw J* (Elsevier) 52:2260–2279
- Atakan B, Akan OB (2010) Deterministic capacity of information flow in molecular nanonetworks. *Nano Commun Netw* 1(1):31–42
- Berg HC (1993) *Random walks in biology*. Princeton University Press, Princeton
- Berg HC, Purcell EM (1977) Physics of chemoreception. *Biophys J* 20(2):193–219
- Brown R (1828) A brief account of microscopical observations. *Philos Mag* 4:161–173
- Bush SF (2010) *Nanoscale communication networks*. Artech House, Boston
- Cao TN, Trinh DP, Jeong Y, Shin H (2015) Anomalous diffusion in molecular communication. *IEEE Commun Lett* 19(10):1674–1677
- Crank J (1975) *The mathematics of diffusion*. Clarendon Press, Oxford, UK
- Culbertson CT, Jacobson SC, Michael Ramsey J (2002) Diffusion coefficient measurements in microfluidic devices. *Talanta* 56(2):365–373
- Delgado J, Alves M, Guedes de Carvalho J (2005) A simple and inexpensive technique to measure molecular diffusion coefficients. *J Phase Equilib Diffus* 26(5):447–451
- Eckford AW (2007) Nanoscale communication with Brownian motion. In: 2007 41st annual conference on information sciences and systems, 14–16 Mar 2007
- Einstein A (1905) On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat. *Ann Phys* 17: 549–560
- Felicetti L, Femminella M, Realì G (2013) Simulation of molecular signaling in blood vessels: software design and application to atherogenesis. *Nano Commun Netw* 4(3):98–119
- Freitas RA (1999) *Nanomedicine, Vol. 1: Basic capabilities*, 1st edn. Landes Bioscience, Austin
- Gohari A, Mirmohseni M, Nasiri-Kenari M (2016) Information theory of molecular communication: directions and challenges. *IEEE Trans Mol Biol Multi-Scale Commun* 2(2):120–142

- Gul E, Atakan B, Akan OB (2010) NanoNS: a nanoscale network simulator framework for molecular communications. *Nano Commun Netw* 1(2):138–156
- Haselmayr W, Aejaz SMH, Asyhari AT, Springer A, Guo W (2017) Transposition errors in diffusion-based mobile molecular communication. *IEEE Commun Lett* 21(9):1973–1976
- Höfling F, Franosch T (2013) Anomalous transport in the crowded world of biological cells. *Rep Prog Phys* 76(4):046602
- IEEE (2016) IEEE recommended practice for nanoscale and molecular communication framework. *IEEE Std* 1906.1-2015, pp 1–64
- Jamali V, Ahmadzadeh A, Wicke W, Noel A, Schober R (2018) Channel modeling for diffusive molecular communication – a tutorial review. In: The proceedings of IEEE (submitted to)
- Kadloor S, Adve RS, Eekford AW (2012) Molecular communication using Brownian motion with drift. *IEEE Trans Nanobioscience* 11(2):89–99
- Koo B, Lee C, Yilmaz HB, Farsad N, Eekford A, Chae C (2016) Molecular MIMO: from theory to prototype. *IEEE J Sel Areas Commun* 34(3):600–614
- Llatser I, Demiray D, Cabellos-Aparicio A, Altılar DT, Alarcón E (2014) N3Sim: simulation framework for diffusion-based molecular communication nanonetworks. *Simul Model Pract Theory* 42:210–222
- Mahfuz MU, Makrakis D, Mouftah HT (2014) A comprehensive study of sampling-based optimum signal detection in concentration-encoded molecular communication. *IEEE Trans Nanobioscience* 13(3):208–222
- Mahfuz MU, Makrakis D, Mouftah HT (2016) Concentration-encoded subdiffusive molecular communication: theory, channel characteristics, and optimum signal detection. *IEEE Trans Nanobioscience* 15(6):533–548
- Mai TC, Egan M, Duong TQ, Renzo MD (2017) Event detection in molecular communication networks with anomalous diffusion. *IEEE Commun Lett* 21(6):1249–1252
- Metzler R, Klafter J (2000) The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Phys Rep* 339(1):1–77
- Nakano T, Okaie Y, Jian-Qin L (2012) Channel model and capacity analysis of molecular communication with Brownian motion. *IEEE Commun Lett* 16(6):797–800
- Nakano T, Eekford AW, Haraguchi T (2013) *Molecular communication*. Cambridge University Press, Cambridge, UK
- Nakano T, Suda T, Okaie Y, Moore M, Vasilakos A (2014) Molecular communication among biological nanomachines: a layered architecture and research issues. *IEEE Trans Nanobioscience* 13:169
- Noel A, Cheung KC, Schober R, Makrakis D, Hafid A (2017) Simulating with AcCoRD: actor-based communication via reaction–diffusion. *Nano Commun Netw* 11:44–75
- Pierobon M, Akyildiz IF (2010) A physical end-to-end model for molecular communication in nanonetworks. *IEEE J Sel Areas Commun* 28(4):602–611

- Pierobon M, Akyildiz IF (2011) Noise analysis in ligand-binding reception for molecular communication in nanonetworks. *IEEE Trans Signal Process* 59(9):4168–4182

Budget Feasibility

► Budget Feasible Mechanisms

Budget Feasible Mechanisms

Hamed Shah-Mansouri and Vincent W. S. Wong
Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada

Synonyms

[Budget feasibility](#); [Combinatorial auctions](#); [Incentive compatibility](#)

Definitions

Budget feasible mechanisms refer to procurement combinatorial auctions in which the auctioneer with a limited budget chooses a subset of items with the goal of maximizing its valuation.

Background

Mobile and wireless systems are evolving at a rapid pace due to the tremendous growth of data traffic. The exceedingly large number of devices and the emergence of new applications and services make the design and management of wireless systems even more challenging. However, traditional models are not capable of supporting future systems such as fifth generation wireless

networks and Internet of Things (Wong et al. 2017).

Auction theory is an applied branch of mathematics and is promising for design, management, and economic analysis of wireless systems. Auction theory is widely applied in different areas of mobile and wireless systems including mobile crowdsensing (e.g., Shah-Mansouri and Wong 2015; Song et al. 2017), spectrum sharing (e.g., Huang et al. 2006; Niyato et al. 2009; Wu and Vaidya 2013), radio resource allocation (e.g., Zhang et al. 2013), multiple access management (e.g., Akkarajitsakul et al. 2011), device-to-device communication (e.g., Xu et al. 2013), and virtualization of fifth generation wireless networks (e.g., Zhu and Hossain 2016).

Auction Mechanisms

Auction is a process to buy and sell commodities and services and is characterized by a set of rules determined by an intermediate agent who manages the auction process. This agent is referred as the auctioneer. Auctions are modeled as interactions among buyers/sellers and the auctioneer. Auction commodities are exchanged based on submitting bids and asks to the auctioneer and receiving payments according to the auction rules. Buyers (sellers) submit their bids (asks) for the commodities and services to the auctioneer. The auctioneer either accepts the bids (asks) or rejects them and clears the market based on the auction rules. There are different types of auction designs. The auction mechanisms can be categorized into the following types based on the role of buyers and sellers.

- *Forward auctions*: The buyers request to obtain the commodities and services from sellers by submitting increasingly higher bids.
- *Reverse auctions*: The sellers compete to offer their commodities and services to buyers by submitting decreasingly lower asks.

Auctions can also be categorized into single-sided and double-sided as follows:

- *Single-sided auctions*: In a single-sided auction, only buyers or sellers submit their bids or asks to the auctioneer.
- *Double-sided auctions*: In a double-sided auction, both buyers and sellers participate in market interactions by submitting their bids and asks.

Other types of auctions are studied by Krishna (2009). Auctions are composed of two major components, which are the winner determination and payment schemes. The winner determination scheme selects a subset of participating players as the winners of the auction. The payment scheme determines the clearing price of commodities and services exchanged during the auction.

Incentive Compatibility

Auction is a powerful tool to model and manage the market of exchanging commodities and services. However, strategic behavior of players may degrade the auction's outcome. Selfish players may misreport the true cost or valuation of their commodities and services in order to gain higher payoffs. For example, a seller may ask for a price higher than the cost associated with its services. Incentive compatibility, which is also referred as truthfulness, is an important property of auctions. An auction mechanism is incentive compatible if reporting the true cost or valuation is the dominant strategy of the players. That is, a player cannot achieve a higher payoff by deviating from the true cost or valuation regardless of other players' strategies. To be immune against the strategic behavior of selfish players, an auction mechanism should be incentive compatible. To achieve incentive compatibility, as shown by Nisan et al. (2007), the winner determination scheme should be monotone, and the payment scheme should be based on critical payments. In a reverse auction, the monotonicity of the winner determination scheme means that if a bidder can win the auction with a certain bid, it can also win the auction with any lower bids. Moreover, the critical payment is the highest bid that the bidder can submit and win the auction.

Utilizing the critical payments is necessary to achieve incentive compatibility. This however may require a large amount of payment. The Vickrey-Clarke-Groves (VCG) mechanism, which is a well-known auction mechanism, may impose an unbounded payment to achieve incentive compatibility (Vickrey 1961; Krishna 2009). This essentially prevents the auctioneer to design a budget feasible mechanism based on VCG. Moreover, the VCG mechanism allocates the items in a socially optimal manner. However, since the budget feasible mechanisms are involved with combinatorial optimization, the VCG mechanism becomes computationally inefficient to manage large markets.

Combinatorial Auctions

Combinatorial auctions refer to a type of market where discrete items are exchanged, and the buyers (sellers) can place bids (asks) on combinations of these items. A survey of combinatorial auctions is provided by De Vries and Vohra (2003). Combinatorial auctions are widely used to model and manage the wireless systems due to the discrete nature of resources and services. Combinatorial auctions are more complex than those auctions which are only involved with commodities of continuous quantity. Due to the computational complexity of combinatorial auctions to achieve an optimal allocation of commodities, several algorithms have been introduced to find sub-optimal allocations. A combinatorial auction is computationally efficient if it can clear the market in polynomial time. In other words, if both winner determination and payment schemes can be terminated in polynomial time, the auction is efficient. Moreover, to quantify how good a sub-optimal allocation, the approximation ratio is investigated. The approximation ratio is the ratio between the optimal outcome (e.g., services' valuation) of the auction and the outcome obtained by an approximate algorithm.

Budget Feasible Mechanisms

Budget feasible mechanisms refer to procurement auctions in which the total payment does

Budget Feasible Mechanisms, Table 1 Comparison of budget feasible mechanisms and their approximation ratios for large-scale systems

Mechanism	Approximation ratio
Singer (2010)	117.7
Chen et al. (2011)	$\frac{5e}{e-1} \approx 7.91$
Song et al. (2017)	$\frac{2e}{e-1} \approx 3.16$

not exceed a limited budget. From now on, this article focuses on a single-sided reverse auction mechanism where the auctioneer with a limited budget chooses a subset of items. The objective of the auctioneer is to maximize its valuation while satisfying budget feasibility. On the other hand, each seller submits a bid in order to maximize its payoff, which is the payment received from the auctioneer minus the cost. Different budget feasible mechanisms have been proposed in the literature as listed in Table 1.

Singer (2010) proposed an incentive compatible budget feasible mechanism for general submodular valuation functions as an important class of set functions. In this mechanism, the auctioneer aims to maximize the valuation function by selecting a subset of items. Since maximizing a submodular function under a budget constraint is known to be NP-hard, a greedy-based approximation method is proposed. To derive the winner determination scheme, a valuation maximization problem is formulated when the payments are substituted by the bids and the auctioneer spends a fixed fraction of the budget. The items are greedily selected based on their marginal contribution to the valuation function in order to achieve a computationally efficient but approximate solution. The payment to each bidder is then determined based on the critical payment. Singer (2010) first showed that such budget feasible mechanism is incentive compatible. He further derived an upper bound on the approximation ratio of the mechanism by showing that for any submodular functions, the ratio between the valuation obtained by an optimal mechanism and that of the proposed greedy-based mechanism is no larger than 117.7.

Chen et al. (2011) further analyzed the aforementioned budget feasible mechanism and pro-

vided a tight bound on the approximation ratio of the greedy-based mechanism. They showed that for any submodular functions, the approximation ratio is less than or equal to $\frac{5e}{e-1} \approx 7.91$.

Song et al. (2017) also proposed an incentive compatible budget feasible mechanism for general submodular functions and applied the mechanism to a mobile crowdsensing system to model the market. Inspired by Singer (2010), a valuation maximization problem is formulated when the payments are substituted by the bids. Singer (2010) used a constant fraction of the budget to determine the winners. However, Song et al. (2017) determined the required fraction of budget based on the market dynamics (e.g., the bids and the valuation function). The required fraction of budget to achieve incentive compatibility essentially depends on the bids submitted by winners, which are not known *a priori*. Song et al. (2017) employed an adaptive approach and iteratively updated the amount of spent budget when determining the winners. They analyzed the approximation ratio of their proposed budget feasible mechanism and showed that it always outperforms the mechanisms proposed by Singer (2010) and Chen et al. (2011). They further showed that when the number of winners tends to infinity, the approximation ratio approaches $\frac{2e}{e-1} \approx 3.16$. This illustrates the applicability of the budget feasible mechanism proposed by Song et al. (2017) for large-scale systems such as mobile crowdsensing.

In addition to the budget feasible mechanisms presented in Table 1, there are several other incentive mechanisms in the literature. Although they achieve incentive compatibility and budget feasibility, they do not provide the approximation ratio of their sub-optimal allocation mechanism in general cases. Anari et al. (2014) analyzed the budget feasible mechanism proposed by Chen et al. (2011) for large-scale crowdsensing systems and proved that the approximation ratio approaches $1 - 1/e$ in this case. Zhao et al. (2016) proposed an online budget feasible mechanism for mobile crowdsensing systems. They further showed that the ratio between the social welfare obtained by their proposed online mechanism and

optimal social welfare of an offline mechanism is constant. Furthermore, Dobzinski et al. (2011) proposed a budget feasible mechanism for more general valuation functions. They showed that for a market with n participants, the approximation ratio of the sub-optimal allocation is $O(\log^2 n)$ for sub-additive valuation functions. Bei et al. (2012) provided a tighter upper bound on the approximation ratio of the mechanism proposed by Dobzinski et al. (2011). Moreover, Chan and Chen (2016) proposed a budget feasible mechanism for a market in which a dealer aims to maximize its revenue when buying items from sellers and selling them to buyers.

Conclusions

Budget feasible mechanisms are promising for the design and management of a broad set of mobile and wireless systems. They are promising to manage realistic markets that need to be immune against the strategic behavior of the users by achieving incentive compatibility but have a limited budget. This article reviewed the existing budget feasible mechanisms and compared their efficiency in achieving close-to-optimal allocations.

References

- Akcarajitsakul K, Hossain E, Niyato D, Kim DI (2011) Game theoretic approaches for multiple access in wireless networks: a survey. *IEEE Commun Surv Tutor* 13(3):372–395
- Anari N, Goel G, Nikzad A (2014) Mechanism design for crowdsourcing: an optimal $1 - 1/e$ competitive budget-feasible mechanism for large markets. In: *Proceedings of IEEE annual symposium on foundations of computer science (FOCS)*, Philadelphia, Oct 2014
- Bei X, Chen N, Gravin N, Lu P (2012) Budget feasible mechanism design: from prior-free to bayesian. In: *Proceedings of annual ACM symposium on theory of computing (STOC)*, New York
- Chan H, Chen J (2016) Budget feasible mechanisms for dealers. In: *Proceedings of ACM international conference on autonomous agents & multiagent systems (AAMAS)*, Singapore
- Chen N, Gravin N, Lu P (2011) On the approximability of budget feasible mechanisms. In: *Proceedings of ACM-*

- SIAM symposium on discrete algorithms (SODA), San Francisco, Jan 2011
- De Vries S, Vohra RV (2003) Combinatorial auctions: a survey. *INFORMS J Comput* 15(3):284–309
- Dobzinski S, Papadimitriou CH, Singer Y (2011) Mechanisms for complement-free procurement. In: *Proceedings of ACM conference on electronic commerce (EC)*, San Jose
- Huang J, Berry RA, Honig ML (2006) Auction-based spectrum sharing. *Mob Netw Appl* 11(3):405–418
- Krishna V (2009) *Auction theory*, 2nd edn. Academic, Burlington
- Nisan N, Roughgarden T, Tardos E, Vazirani VV (2007) *Algorithmic game theory*. Cambridge University Press, Cambridge
- Niyato D, Hossain E, Han Z (2009) Dynamic spectrum access in IEEE 802.22-based cognitive wireless networks: a game theoretic model for competitive spectrum bidding and pricing. *IEEE Wirel Commun* 16(2):16–23
- Shah-Mansouri H, Wong VWS (2015) Profit maximization in mobile crowdsourcing: a truthful auction mechanism. In: *Proceedings of IEEE ICC*, London, June 2015
- Singer Y (2010) Budget feasible mechanisms. In: *Proceedings of IEEE symposium on foundations of computer science (FOCS)*, Las Vegas, Oct 2010
- Song B, Shah-Mansouri H, Wong VWS (2017) Quality of sensing aware budget feasible mechanism for mobile crowdsensing. *IEEE Trans Wirel Commun* 16(6):3619–3631
- Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. *J Financ* 16(1):8–37
- Wong VWS, Schober R, Ng DWK, Wang LC (2017) *Key technologies for 5G wireless systems*. Cambridge University Press, West Nyack
- Wu F, Vaidya N (2013) A strategy-proof radio spectrum auction mechanism in noncooperative wireless networks. *IEEE Trans Mob Comput* 12(5):885–894
- Xu C, Song L, Han Z, Zhao Q, Wang X, Cheng X, Jiao B (2013) Efficiency resource allocation for device-to-device underlay communication systems: a reverse iterative combinatorial auction based approach. *IEEE J Sel Areas Commun* 31(9):348–358
- Zhang Y, Lee C, Niyato D, Wang P (2013) Auction approaches for resource allocation in wireless systems: a survey. *IEEE Commun Surv Tutor* 15(3):1020–1041
- Zhao D, Li XY, Ma H (2016) Budget-feasible online incentive mechanisms for crowdsourcing tasks truthfully. *IEEE/ACM Trans Netw* 24(2):647–661
- Zhu K, Hossain E (2016) Virtualization of 5G cellular networks as a hierarchical combinatorial auction. *IEEE Trans Mob Comput* 15(10):2640–2654

Byzantine Attack and Defense in Wireless Data Fusion

Linyuan Zhang¹, Guoru Ding², and Qihui Wu³

¹Jiangnan Institute of Computing Technology, Wuxi, China

²College of Communications Engineering, Army Engineering University of PLA, Nanjing, China

³Department of Electronics and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Synonyms

Data falsification in wireless data fusion; Distributed inference with wireless Byzantine data

Definition

Byzantine attack in wireless data fusion means that part of the data from malicious Byzantine attackers may be fake, which makes the results of data fusion deviate from the truth, while Byzantine defense, on the contrast, is the process to decrease negative effects of Byzantine attack on the data fusion via various defense methods.

Historical Background

Byzantine attack stems from the Byzantine generals problem in Lamport et al. (1982) as follows:

A group of Generals of the Byzantine army camped with their troops around an enemy city. Communicating only by messenger, the generals must agree upon a common battle plan. However, one or more of them may be traitors who will try to confuse the others. The problem is to find an algorithm to ensure that the loyal generals will reach agreement.

Based on the illustration above, it is well acknowledged that Byzantine attack exists in various distributed sensor networks (Varshney

1996), e.g., cooperative spectrum sensing in cognitive radio networks (Zhang et al. 2015) and collaborative state estimation in smart grids (Kosut et al. 2011).

In scenarios of wireless data fusion, Byzantine attack is a kind of insider attack in PHY layer, and it refers to the falsification of wireless data (Vempaty et al. 2013). Specifically, in the process of wireless data fusion to make cognition about the wireless environment, lots of devices are equipped and volumes of data are produced. Due to various factors, such as device failure or malicious invasion, part of data sources become Byzantine attackers where data are falsified. As a result, fake data are intermixed with real data, which brings about deviation of data fusion results. In particular, in Marano et al. (2009), Rawat et al. (2011), and Kailkhura et al. (2015), the optimal attack strategies are analyzed, which maximize the detection error probability, and it is pointed out that when Byzantine attackers occupy a certain fraction of all sensors, data fusion scheme with desired performance becomes completely incapable. Hence, robust data fusion is quite challenging in the presence of Byzantine attack. In particular, Chen et al. (2008) is one of the first studies that identifies the problem of Byzantine attack, also known as spectrum sensing data falsification (SSDF) attack, in the context of cooperative spectrum sensing in cognitive radio networks. Further, Ding et al. (2013, 2014) exploit machine learning methods to eliminate fake data, including kernel-based learning and data cleaning methods. During the past decades, Byzantine attack and defense in wireless data fusion have received increasing attentions, especially with the quick development of software-defined radios (SDR). On the one hand, the reconfigurability of SDR enables devices to behave more autonomously, and devices are vulnerable to be invaded and controlled by attackers. On the other hand, the wireless data is closely related to benefits, e.g., spectrum access opportunities, which become the motivation of making data falsification.

Foundations

Wireless data fusion is the process of integrating multiple data sources to achieve more consistent, accurate, and useful information about wireless signals than that provided by any individual data source. Considering the high uncertainty brought by the complicated wireless environment and the special wireless signal propagation characteristic, wireless data fusion is necessary to break the limitation of a single data source through exploiting diversity gains of multiple data sources. Nevertheless, it is built on the basic assumption that all data sources are honest, i.e., the data really reflects the received signal's characteristics. In fact, due to the openness of wireless channels, distributed wireless networks are prone to various security threats, e.g., jamming, spoofing, and wiretap, among which data falsification, i.e., Byzantine attack, breaks the assumption above. In particular, it destroys the inherent relation between the data and the wireless target and deteriorates the reliability of data fusion.

In the past years, Byzantine attack and defense in cooperative spectrum sensing (CSS) have received wide attention, which can be taken as a classic example in wireless data fusion. First, cognitive radio opens a promising paradigm to improve radio spectrum utilization by allowing cognitive or unlicensed secondary users (SUs) to intelligently employ spectrum holes unused by licensed primary users (PUs), where spectrum sensing is one of key enabling techniques to detect spectrum holes. Then, to combat wireless channel impairments, such as multipath fading and shadowing, CSS has been proposed as an effective method to improve the sensing performance by exploiting the spatial diversity among multiple spectrum sensors, but at the same time, Byzantine attack is introduced. Here, the main goals of Byzantine attackers in CSS are twofold: the first is to decrease the detection probability for disturbing the normal operation of PU systems, and the second is to increase the probability of false alarm with the purpose of depriving access opportunities of the honest SUs.

Its attack performance is closely related to several aspects: attack scenario, attack basis, attack opportunity, and attack population (Zhang et al. 2015). Specifically,

- Attack scenario depicts where attackers act in, i.e., the network circumstance. It is classified into centralized and decentralized based on whether there is a dedicated node collecting all sensors' data and making data fusion. It determines the way of sensors cooperating, greatly influencing attack behaviors (Yan et al. 2012), and is chosen as the classification standard of networks.
- Attack basis denotes information based on which malicious users falsify measurements and solves how to launch attacks. Attack basis include local observation, other sensors' sensing results, fusion rules, and so on. An abundant attack basis can assist malicious users to design effective attack strategies.
- Attack opportunity answers about when to attack. At a certain slot, malicious users may decide whether or not to attack with a certain probability. Besides, attackers may take into consideration of this current state, such as sensing results and attack expectation. Attack opportunity reveals the flexibility and variety of attack behaviors.
- Attack population is a basic metric of the Byzantine attack answering about who to attack, and is related to the percentage of malicious sensors in all sensors, dedicating to what degree of severity networks suffers from attackers. Naturally, as the attack population increases, results of data fusion lean more to the Byzantine attacks, and when Byzantine attackers are a certain fraction of all sensors, data fusion scheme becomes completely incapable (Kailkhura et al. 2015).
- Data fusion results. Byzantine data makes the data fusion results biased to certain extend, but when the attack is not very strong, i.e., attack population is lower than 0.5, the bias is not large and can still be used to evaluate the reliability of every data resources (Rawat et al. 2011).
- Trusted data resources' assistance. In certain cases, it is assumed that certain honest sensor is known to the fusion center and its report is approximated as the ground truth of the spectrum state and is used to evaluate other sensors' reports.
- Side information about the wireless environment, including the channel propagation characteristics, e.g., showing correlation (Min et al. 2011), Markovian model for the spectrum state (He et al. 2013), and so on. Side information can be used to extract related parameters of reported data and evaluate the matching degree, based on which the reliability evaluation is done.
- Stale information about real states of the target. The stale information is used to denote information that can reflect the real channel states but is overdue for spectrum sensing in current slot, such as the transmit results, and the whole-slot sensing (Zhang et al. 2016). Although the stale information cannot help spectrum access in current slot, it can assist identification of Byzantine attackers. Its main advantage is that its reliability is rarely affected by the attack behaviors. That means that it can work in severe environments, e.g., Byzantine attackers are in majority.

Further, to exploit all sensors' data instead of abandoning suspicious sensors' data, Ding et al. (2014) proposes a data cleansing-based robust cooperative sensing scheme, where the underutilization of licensed spectrum bands and the sparsity of nonzero abnormal data are jointly exploited to robustly cleanse out the nonzero abnormal data component, not the sensing data itself, from the original corrupted sensing data.

Finally, the Byzantine attack problem is not only a threat in terms of data security but also a

In Byzantine defense, one key problem is how to identify Byzantine attacker, and an effective approach is to build a reference to evaluate the reporters' reliability, which includes but is not limited to the following:

game between attackers and the detection system. On one side, false sensing outputs from malicious alteration ruin the integrity of data and lower the reliability of data fusion. On the other side, like spear and shield, the Byzantine attack and defense are mutually exclusive but indivisible. When launching the attack, the attackers tend to maximize their attack gains, at the same time of considering how to escape from being found by the defense system, while in turn the diversity and flexibility of attack strategies force the defense system to enhance its universality and efficiency. All of these make the issue of Byzantine attack and defense very challenging and quite complicated.

Key Applications

Byzantine attack and defense in wireless data fusion are an important issue of various distributed sensor networks.

Cross-References

- [Anomaly Detection for IoT Systems](#)
- [Application of Machine Learning in Wireless Sensor Network](#)
- [Cognitive Radio Sensor Networks](#)
- [Data Gathering in Wireless Sensor Networks](#)
- [Pilot Spoofing Attack](#)

References

- Chen R, Park JM, Bian K (2008) Robust distributed spectrum sensing in cognitive radio networks. In: IEEE INFOCOM 2008 – the 27th conference on computer communications
- Ding G, Wu Q, Yao YD, Wang J, Chen Y (2013) Kernel-based learning for statistical signal processing in cognitive radio networks: theoretical foundations, example applications, and future directions. *IEEE Signal Process Mag* 30(4):126–136
- Ding G, Wang J, Wu Q, Zhang L, Zou Y, Yao YD, Chen Y (2014) Robust spectrum sensing with crowd sensors. *IEEE Trans Commun* 62(9):3129–3143
- He X, Dai H, Ning P (2013) A Byzantine attack defender in cognitive radio networks: the conditional frequency check. *IEEE Trans Wirel Commun* 12(5):2512–2523
- Kailkhura B, Han YS, Brahma S, Varshney PK (2015) Distributed Bayesian detection in the presence of Byzantine data. *IEEE Trans Signal Process* 63(19):5250–5263
- Kosut O, Jia L, Thomas RJ, Tong L (2011) Malicious data attacks on the smart grid. *IEEE Trans Smart Grid* 2(4):645–658
- Lamport L, Shostak R, Pease M (1982) The Byzantine generals problem. *ACM Trans Program Lang Syst* 4(3):382–401
- Marano S, Matta V, Tong L (2009) Distributed detection in the presence of Byzantine attacks. *IEEE Trans Signal Process* 57(1):16–29
- Min AW, Shin KG, Hu X (2011) Secure cooperative sensing in IEEE 802.22 WRNs using shadow fading correlation. *IEEE Trans Mobile Comput* 10(10):1434–1447
- Rawat AS, Anand P, Chen H, Varshney PK (2011) Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks. *IEEE Trans Signal Process* 59(2):774–786
- Varshney PK (1996) Distributed detection and data fusion. Springer, New York
- Vempaty A, Tong L, Varshney PK (2013) Distributed inference with Byzantine data: state-of-the-art review on data falsification attacks. *IEEE Signal Process Mag* 30(5):65–75
- Yan Q, Li M, Jiang T, Lou W, Hou YT (2012) Vulnerability and protection for distributed consensus-based spectrum sensing in cognitive radio networks. In: 2012 Proceedings IEEE INFOCOM, pp 900–908
- Zhang L, Ding G, Wu Q, Zou Y, Han Z, Wang J (2015) Byzantine attack and defense in cognitive radio networks: a survey. *IEEE Commun Surv Tutor* 17(3):1342–1363
- Zhang L, Ding G, Wu Q, Song F (2016) Defending against Byzantine attack in cooperative spectrum sensing: defense reference and performance analysis. *IEEE Access* 4:4011–4024

Cache Placement and Content Delivery Design in Wireless Caching Networks

Ruijin Sun
Pengcheng Laboratory, Shenzhen, China

Synonyms

Caching strategy; Content transmission strategy

Definition

Cache placement and content delivery design is to allocate the storage resource and the communication resource in wireless networks to achieve better system performance.

Historical Background

With the widespread popularity of smart mobile devices (tablets, pads, phones, etc.) and the development of mobile Internet, the wireless traffic data is increasing explosively. It has been reported in I. Cisco Visual Networking (2017) that the mobile data traffic will grow sevenfold from 2016 to 2021 and will reach 49 exabytes per month by 2021. Meanwhile, the wireless services have been shifted from the traditional voices and messages to the data-craving services, such as

video streaming, online games, augmented reality (AR), virtual reality (VR), and so on. Among them, the video streaming has been the dominant source of the wireless mobile traffic and will make up 78% of the total traffic by the end of 2021 (I. Cisco Visual Networking 2017).

To keep pace with this unprecedented traffic requirement, great efforts have been made to expand the network capacity from both the network architecture evolution and the wireless transmission technologies. For the network architecture evolution, ultra-dense networks are proposed to significantly shrink the cell size for the spectrum reuse. Besides, cloud radio access network (Cloud-RAN) architecture is also put forward to process baseband signals of all base stations (BSs) in a central processor to enhance the system rate. For the wireless transmission technology, massive multiple input and multiple output (Massive MIMO), millimeter wave (mmWave), and visible light communication (VLC) are proposed to improve the network capacity (Wong et al. 2017). All these technologies can boost the peak rate of the network; however, they will lead to high infrastructure cost and heavy backhaul traffic burden.

Wireless edge caching is proposed to deal with this issue from the service transmission characteristics, especially for the video traffic feature. More specifically, 80% of video traffic is caused by 20% popular videos, and the user request probability for the video streaming can be precisely predicted. Therefore, pre-storing popular contents at BSs or mobile devices during off-peak

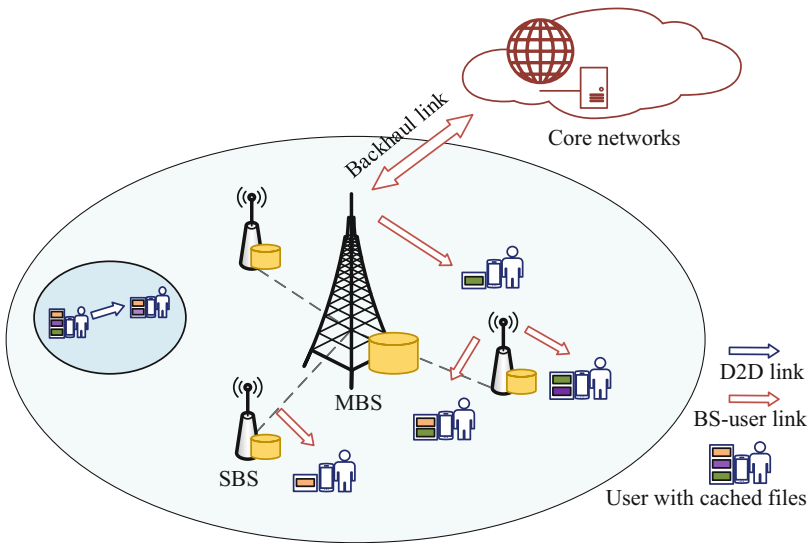
hours can avoid the redundant transmission in core networks and hence significantly relieve the backhaul traffic burden. As demonstrated in Patel et al. (2014), pre-caching popular contents at BSs can reduce the backhaul traffic to 35%. Moreover, with such local caching, contents are brought much closer to users and thus the content access delay can be considerably reduced. Based on the location of the content caching, wireless caching networks can be classified into two categories: caching at BSs and caching at mobile users. Compared with the BS caching, caching at mobile users with the aid of device-to-device (D2D) communications can further offload the BS traffic to D2D users.

Generally, wireless caching consists of two stages: the caching stage and the delivery stage. In the caching stage, popular contents are stored by BSs or mobile users according to the cache placement algorithm. Then, in the delivery stage, the cached contents can be directly transmitted by BSs or mobile users, which can significantly relieve the backhaul pressure and reduce the end-to-end content delivery delay. As a result, there are two basic issues in wireless caching networks. The first issue is how to reasonably design cache placement algorithm in the caching stage. Different from wired networks, the caching strategy

in wireless networks is more complicated due to the broadcast nature of the wireless channel and the user mobility. The second issue is how to rationally design the content transmission strategy with the given caching status in the delivery stage. Different from traditional wireless networks, the transmission strategy in wireless caching networks should not only consider the channel conditions but also the cache status of BSs and the user requests.

Foundations

The wireless caching network architecture is shown in Fig. 1, which consists of core networks, macro BS (MBS), small BSs (SBSs), and mobile users. The MBS is connected to core networks via the backhaul link, and each SBS is also connected to MBS via the backhaul link. The core networks can access to all contents in the content library. Both BSs (MBS and SBSs) and mobile users are equipped with storage spaces to cache popular contents for the backhaul capacity consumption minimization and the end-to-end delay reduction. Based on this architecture, when a mobile user requests a content, it first searches the content from its own storage. If the requested



Cache Placement and Content Delivery Design in Wireless Caching Networks, Fig. 1 The wireless caching network architecture

content is not cached by the user itself, the user will be served by nearby users via the D2D communication or the BS. If the desired content cannot even be found at BSs, BSs will first fetch the content from core networks and then transmit the content to the user.

Clearly, the caching performance is highly related to the content popularity, which is usually obtained for a certain region during a period of time. According to Breslau et al. (1999), the popularity of all contents in a particular library follows a Zipf distribution. In addition to the general Zipf distribution, the content popularity can also be predicted by analyzing the huge historical content requests and using the machine learning (ML)-based prediction methods. In this way, the specific traffic features and network characteristics for different scenarios can be fully mined. Generally, the content popularity changes much slower than the traffic variation and channel fading of wireless networks. For instance, a hot film may last for one week, while the wireless channel fading always changes at the millisecond level. Therefore, the content popularity is regarded as a constant for a long time.

With the given content popularity, the cache placement and content delivery strategy are designed to optimize the system performance. A commonly used caching performance metric is the cache hit ratio, which is defined as a ratio of the number of user requests being satisfied by local cached contents to the total number of user requests. Except for the cache-based metric, the metric of wireless networks should also be considered. As the pre-caching can effectively relieve the backhaul pressure and considerably reduce the end-to-end delay, the backhaul capacity consumption and the content delivery delay are important metrics to evaluate the caching performance. Besides, traditional system metrics, such as the network capacity, the energy efficiency, the transmission power, the quality of user experience (QoE), and so on, can also be adopted. In the following, we will detail the cache placement and the content delivery design to optimize these caching performance metrics.

Cache Placement Design

Different from wired networks, the caching strategy in wireless networks is more complicated due to the broadcast nature of the wireless channel and the user mobility. Intuitively, it seems that the most popular caching strategy, which stores the most popular contents until the cache capacity is full, can achieve the best system performance, such as the cache hit ratio maximization, the backhaul traffic minimization, the end-to-end delay reduction, and so on. This is true for the single-BS network (Gong et al. 2017). However, for a multi-BS network where users are cooperatively served by all BSs, caching strategies of these BSs will be mutually influenced and should be redesigned considering the user association. With the fixed network topology, the content caching strategies of multiple cache-enabled helpers are jointly optimized to minimize the average content delivery time (Shanmugam et al. 2013). Then, from the perspective of spatial stochastic networks, the caching strategies of multiple BSs are also studied to reduce the average network latency (Zhang et al. 2018).

Furthermore, the user mobility feature also plays a key role in the caching strategy design, especially for high-mobility scenario, e.g., highway communication networks. By modeling the user movements that moving users associate with different BSs as a Markov chain, the caching strategies of multiple BSs are designed to minimize the backhaul traffic load (Poularakis and Tassiulas 2017). Then, for D2D caching networks, an inter-contact model, where each user pair has two states based on the communication region, i.e., the contact state and the inter-contact state, is adopted to characterize the user mobility (Passarella and Conti 2013). Based on this model, mobility-aware caching strategies at mobile users are optimized to maximize the D2D data offload ratio (Wang et al. 2017) and minimize the average transmission cost (Deng et al. 2018), respectively.

Content Delivery Design

With the given caching status, wireless network resources need to be rationally assigned for the content delivery. Different from traditional wire-

less networks, the transmission strategy in wireless caching networks should not only consider the channel conditions but also the cache status of BSs and the user requests. For instance, a user may prefer to associate with a BS caching the requested content to minimize the backhaul traffic load. In Park et al. (2016), for a cloud-RAN scenario where multiple users are cooperatively served by multiple BSs, beamforming vectors of multiple BSs are jointly optimized to improve the system rate. Since different users have a high chance to request the same content, especially when the content popularity distribution is concentrated, the multicast transmission is adopted to minimize the weighted sum of the backhaul traffic load and the transmission power (Tao et al. 2016) and to maximize the weighted sum of users' QoE (Sun et al. 2018). Clearly, the cache placement in the caching stage has a significant impact on the content delivery design. With a reasonable caching strategy, the transmission strategy can be improved, and the system performance can be enhanced.

Key Applications

As pre-caching can effectively relieve the backhaul traffic load and considerably reduce the end-to-end content delivery delay, there is a great interest motivated by profit to deploy wireless caching networks in practice. The network operator can provide cache servers for BSs, and the content provider can lease the storage space of these cache servers to cache popular contents according to caching strategies. Then, a majority of user requests can be directly served by BSs without the backhaul capacity consumption. Note that the deployment of wireless caching networks has no significant infrastructure change and is very cost-effective.

Cross-References

- [Caching Strategy](#)
- [Content Delivery](#)

References

- Breslau L, Cao P, Fan L, Phillips G, Shenker S (1999) Web caching and Zipf-like distributions: evidence and implications. In: Proceedings of IEEE international conference on computer communications (INFOCOM), New York, Mar 1999, pp 126–134
- Deng T, Ahani G, Fan P, Yuan D (2018) Cost-optimal caching for D2D networks with user mobility: modeling, analysis, and computational approaches. *IEEE Trans Wirel Commun* 17(5):3082–3094
- Gong J, Zhou S, Zhou Z, Niu Z (2017) Policy optimization for content push via energy harvesting small cells in heterogeneous networks. *IEEE Trans Wirel Commun* 16(2):717–729
- I. Cisco Visual Networking (2017) Global mobile data traffic forecast update, 2016–2021 white paper, Feb 2017. <http://goo.gl/yITuVx>
- Park S-H, Simeone O, Shitz SS (2016) Joint optimization of cloud and edge processing for fog radio access networks. *IEEE Trans Wirel Commun* 15(11):7621–7632
- Passarella A, Conti M (2013) Analysis of individual pair and aggregate intercontact times in heterogeneous opportunistic networks. *IEEE Trans Mob Comput* 12(12):2483–2495
- Patel M, Naughton B, Chan C, Sprecher N, Abeta S, Neal A et al (2014) Mobile-edge computing introductory technical white paper. White Paper, Mobile-edge Computing (MEC) industry initiative, Sept 2014
- Poularakis K, Tassioulas L (2017) Code, cache and deliver on the move: a novel caching paradigm in hyperdense small-cell networks. *IEEE Trans Mob Comput* 16(3):675–687
- Shanmugam K, Golrezaei N, Dimakis AG, Molisch AF, Caire G (2013) Femtocaching: wireless content delivery through distributed caching helpers. *IEEE Trans Inf Theory* 59(12):8402–8413
- Sun R, Wang Y, Cheng N, Zhou H, Shen X (2018) QoE driven BS clustering and multicast beamforming in cache-enabled C-RANs. In: Proceedings of IEEE international conference on communications (ICC), Kansas City, Oct 2018, pp 1–6
- Tao M, Chen E, Zhou H, Yu W (2016) Content-centric sparse multicast beamforming for cache-enabled cloud RAN. *IEEE Trans Wirel Commun* 15(9):6118–6131
- Wang R, Zhang J, Song S, Letaief KB (2017) Mobility-aware caching in D2D networks. *IEEE Trans Wirel Commun* 16(8):5001–5015
- Wong VWS, Schober R, Ng DWK, Wang LC (2017) Key technologies for 5G wireless systems. Cambridge University Press, Cambridge, UK
- Zhang S, He P, Suto K, Yang P, Zhao L, Shen X (2018) Cooperative edge caching in user-centric clustered mobile networks. *IEEE Trans Mob Comput* 17:1791–1805

Caching and Computing

- [Integrated System of Networking, Caching, and Computing](#)

Caching Strategy

- [Cache Placement and Content Delivery Design in Wireless Caching Networks](#)

Call Assignment

- [Network Selection in Heterogeneous Wireless Networks](#)

Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space

Lin Cai^{1,2}, Xuemin (Sherman) Shen³, and Jon W. Mark⁴

¹The Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

²E&CE Department, Illinois Institute of Technology, Chicago, IL, USA

³Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

⁴E&CE Department, University of Waterloo, Waterloo, ON, Canada

Synonyms

[Network capacity](#); [Network maximum throughput](#); [Network transport capacity](#)

Definition

Network capacity is defined as the total throughput (bps) of all flows in the network;

Network transport capacity is defined as the sum of the products of each flow's throughput and its transceiver distance in the network ($\text{bit} \cdot \text{m/s}$).

Historic Background

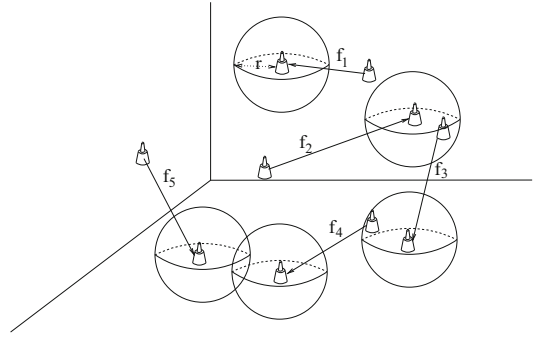
The capacity of point-to-point wireless communication over an additive white Gaussian channel was laid out by Claude Shannon in his 1948 seminal paper (Shannon 1948). With the state-of-the-art channel coding, e.g., turbo code, energy and bandwidth efficiency of practical point-to-point communications is pushing closer to the Shannon limit. Due to the broadcast property of the wireless channel, wireless capacity in a multipoint-to-multipoint *networked* environment is interference limited. How to efficiently deploy the spatial utilization of wireless channel to improve the capacity of wireless networks has been a key challenging issue for several decades. In addition to the time and frequency domains, there is the spatial domain that offers diversity/multiplexing gain. Also, the wireless network capacity depends on network topology. Although asymptotic bounds of wireless network capacity have been heavily pursued, how to fully explore the spatial multiplexing gain of wireless networks and determine the capacity of the networks with random topology are still open issues.

To explore the spatial capacity of wireless networks, multiple flows should be scheduled for concurrent transmissions if the transmitters are sufficiently far away from the non-intended receivers. Given a three-dimensional space and a number of active users in the network, what are the expected number of concurrent transmissions, the network capacity, and the transport capacity of networks? Here, the network capacity is defined as the total throughput (bps) of all flows in the network, and the network transport capacity is defined as the sum of the products of each flow's throughput and its transceiver distance in the network ($\text{bit} \cdot \text{m/s}$). In general, the network capacity is the main performance index for single-hop wireless access networks, while the network transport capacity is for multi-hop wireless networks. The analyti-

cal results obtained reveal what the main factors that affect network (transport) capacity are, which will provide important guidelines for wireless network planning, protocol, and architecture design and optimization.

Capacity bounds of wireless networks have been extensively studied in the literature. In the pioneer work of Gupta and Kumar (2000), the asymptotic bounds of network transport capacity have been derived, given the node density in an arbitrary or random wireless network. They extended their work to three-dimensional space in Gupta and Kumar (2001), considering all nodes are located in a sphere. Since then, a large number of follow-up papers have appeared. The impact of user mobility on network capacity has been studied in Grossglauser and Tse (2001) and Jafar (2005). It was found that node mobility can be exploited to increase the network capacity (Grossglauser and Tse 2001), but excessive mobility contrarily limits the capacity (Jafar 2005). The capacity of different networks with different traffic patterns, e.g., relay traffic, convergent traffic, and broadcast traffic, have been investigated in Gastpar and Vetterli (2002), Gamal (2003), and Alireza et al. (2006). These works studied the capacity region, or the upper/lower bounds on network capacity based on the protocol model and the physical model in Gupta and Kumar (2000), assuming that the transmission rate is a binary function with no rate adaptation in the physical layer. In addition, the asymptotic capacity bounds derived in these works may be too loose to be useful in realistic networks where the network topology is random and may change from time to time due to user mobility.

On the other hand, existing theoretical studies indicate that concurrent transmissions should be allowed to improve the resource utilization, as long as all interferers are outside a well-defined exclusive region (ER) around the destinations (Radunovic and Le Boudec 2004; Cai et al. 2009). It is shown in Cai et al. (2010) and Liu et al. (2008) that ER-based resource management can significantly improve the network throughput by exploiting the spatial dimension of the wireless channel. How to control the interference level



Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space, Fig. 1
Exclusive region for concurrent transmissions

and optimize the exclusive region size to maximize the network capacity and transport capacity of a random network in three-dimensional space are still an open issue.

Key Applications

According to the Shannon limit of an additive white Gaussian noise (AWGN) channel, the upper bound of the achievable data rate of node j is

$$R_j = W$$

$$\log_2 \left(1 + \frac{G_{ij} P_i}{(N_0 W + \sum_{k \in \tau, k \neq i} C_0 G_{kj} P_k)} \right) \text{ bps},$$

where P_i and P_k are the transmission power levels of the sender i and an interferer k , respectively, G_{ij} is the channel gain between nodes i and j , τ is the set of flows that transmit concurrently, C_0 is the cross-correlation coefficient among concurrent transmissions, N_0 is the one-sided power spectral density of additive white Gaussian noise, and W is the signal bandwidth.

In a three-dimensional space, an exclusive region can be a sphere centered at the receiver with radius r , as shown in Fig. 1. In the figure, flows f_1 , f_2 , f_4 , and f_5 can concurrently transmit since all senders are outside the exclusive regions of other receivers; flow f_3 interferes with f_2 or f_4 , and thus they cannot be transmitted concurrently. Different from the guard zone specified in

the protocol model in Gupta and Kumar (2000), the general exclusive region employed in our system model is independent of the transceiver distance but is related to the interference and noise levels.

Without loss of generality, the flows being checked by the scheduling algorithm are labeled by flow 1, 2, ..., N . Given the exclusive region radius r , let $p(k, n)$ denote the probability that k flows satisfy the concurrent transmission condition, after checking the first n ($\leq N$) flows one by one. The first flow f_1 is always scheduled for transmission and added to the set γ and $p(1, 1) = 1$. The second flow f_2 will be added to the set γ if it does not conflict with f_1 . Let the random variable Z be the distance between two devices randomly located in an $l \times l \times h$ three-dimensional (3D) space. Denote Q the probability of another sender outside the exclusive region of a receiver, *i.e.*, the distance between the sender (of flow j) and the receiver (of flow i) is larger than r , $Q = \Pr(d_{j,i} \geq r) = \int_r^{\sqrt{2l^2+h^2}} f_Z(z)dz$, where $f_Z(z)$ is the probability density functions (pdf) of Z (Cai et al. 2009).

Given that the transmitters and receivers are randomly and independently deployed, for flows that can transmit concurrently, the locations of

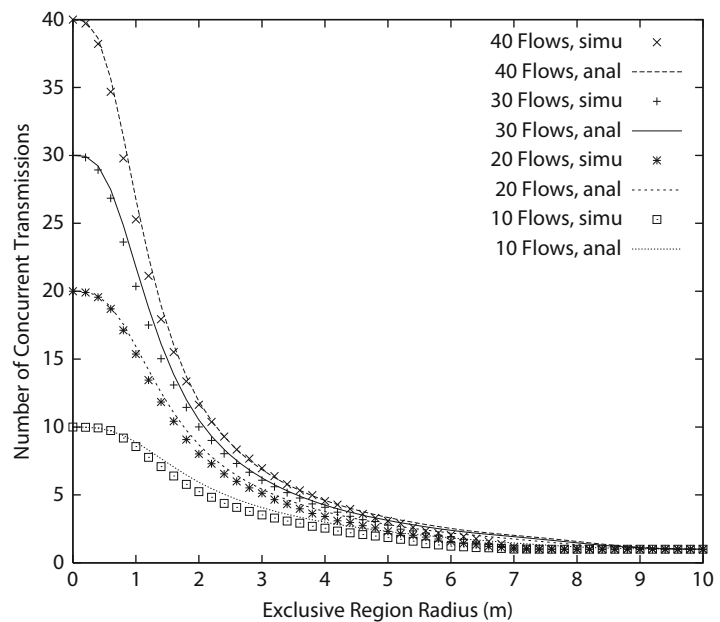
their receivers are independent to each other. Consequently, given that the transmitter of a flow is outside the ER of the receiver of another flow with probability Q , the probability that two transmitters are outside the ERs of two receivers is Q^2 . Thus, the probability that two flows do not conflict with each other is Q^2 , as each sender should be outside the exclusive region of the other flows' receiver. Accordingly, the probability that two flows conflict with each other is $1 - Q^2$. Therefore, $p(2, 2) = Q^2$ and $p(1, 2) = 1 - Q^2$.

By checking the first n flows, there are k flows in the set γ if (a) there are $k - 1$ flows in γ for the first $n - 1$ flows and the n -th flow does not conflict with the $k - 1$ flows in γ or (b) there are k flows in γ after checking the first $n - 1$ flows, and the n -th flow conflicts with at least one of the k flows.

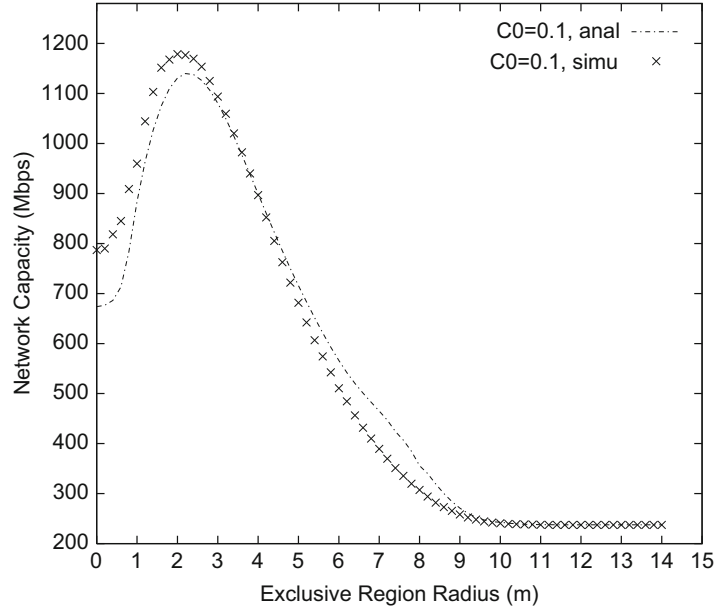
$$p(k, n) = p(k - 1, n - 1)Q^{2(k-1)} + p(k, n - 1)(1 - Q^{2k}) \quad \text{for } k < n. \quad (1)$$

If only the first flow can be added to γ , implying that the following $n - 1$ flows do not satisfy the concurrent transmission condition, $p(1, n) = (1 - Q^2)^{n-1}$. Another extreme case is that all n flows can transmit concurrently, which means

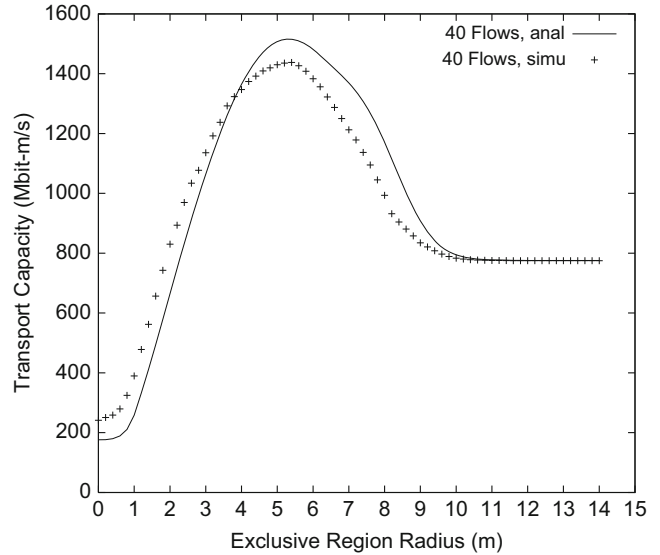
Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space, Fig. 2 Expected number of concurrent transmissions vs ER radius



Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space, Fig. 3 Network capacity vs ER radius ($\alpha = 4$)



Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space, Fig. 4 Transport capacity vs ER radius ($\alpha = 4$)



that none of the flows conflicts with the remaining $n - 1$ flows $p(n, n) = Q^{(n-1)n}$. Given $p(1, 1)$, $p(1, 2)$, and $p(2, 2)$, $p(k, N)$ can be iteratively obtained as a function of r for $1 \leq k \leq N$. The expected number of concurrent transmissions is thus given by

$$E[N] = \sum_{k=1}^N k p(k, N). \quad (2)$$

Given the transmission distance z' , transmission power P , and the channel gain s between the transmitter and the receiver, the received signal power, P_R , is given by $P_R = K G_T G_R z'^{-\alpha} s P$, where K is a constant G_T and G_R are the antenna gains of the transmitter and receiver, respectively. Accordingly, the achievable data rate of the flow is calculated as $R = \eta W \log_2(1 + \frac{K G_T G_R z'^{-\alpha} s P}{N_0 W + I})$, where η is a constant coefficient related to the efficiency of the transceiver design.

Let the random variable Z' denote the effective transmission distance, which can be represented as a truncated random variable over $[d_{\min}, \sqrt{2l^2 + h^2}]$.

Denote the interference distance and the gain of the channel between the i th interferer to a receiver as random variables V_i over $[r, \sqrt{2l^2 + h^2}]$ and S_i , respectively. The pdf of V_i is given by $f_{V_i}(v) = \frac{f_Z(v)}{\int_r^{\sqrt{2l^2 + h^2}} f_Z(z) dz}$. Given the interference distance v and the channel fading factor s_i , the interference from a single interferer is $I_1 = KC_0G_TG_Rv^{-\alpha}s_iP$.

Given that there are k concurrently transmitting flows, each flow has $k - 1$ interferers. Since the received signal strength and the interference strength are independent R.V.s, the received $SINR$ of a single flow can be obtained as

$$SINR|k \approx KG_TG_RPs_z'^{-\alpha} / [N_0W + (k - 1)I_1]. \quad (3)$$

Accordingly, the average achievable transmission rate and the bit-meter product of a single flow under $k - 1$ interferers are given by

$$E[T_S|k] \approx \iiint \eta W \log_2(1 + \frac{KG_TG_Rz'^{-\alpha}sP}{N_0W + (k - 1)KC_0G_TG_RPv_i^{-\alpha}s_i}) \cdot f_{V_i}(v_i) f_{S_i}(s_i) f_S(s) f_{Z'}(z') dv_i ds_i ds dz', \quad (4)$$

$$E[Tr_S|k] \approx \iiint z' \eta W \log_2(1 + \frac{KG_TG_Rz'^{-\alpha}sP}{N_0W + (k - 1)KC_0G_TG_RPv_i^{-\alpha}s_i}) \cdot f_{V_i}(v_i) f_{S_i}(s_i) f_S(s) f_{Z'}(z') dv_i ds_i ds dz'. \quad (5)$$

Removing the condition of $k - 1$ interferers, $E[T_S]$ and $E[Tr_S]$ can be obtained as $E[T_S] = \sum_{k=1}^N E[T_S|k]p(k, N)$, $E[Tr_S] = \sum_{k=1}^N E[Tr_S|k]p(k, N)$.

The capacity and transport capacity of the network are the sum of the throughput and bit-meter products of all flows concurrently transmitting, which are obtained as $E[T] = \sum_{k=1}^N kE[T_S|k]p(k, N)$, and $E[Tr] = \sum_{k=1}^N kE[Tr_S|k]p(k, N)$.

The expected number of concurrent transmissions, $E[N]$, under various exclusive region sizes is shown in Fig. 2. It can be seen that the expected number of concurrent transmissions decreases when the exclusive region radius r increases.

The network capacity and transport capacity under various ER are shown in Figs. 3 and 4. The experiment parameters are listed in Table 1. It can be seen that the network capacity is bounded by high interference level when r is small; and the capacity becomes constant and equals the average

throughput of a single flow with a TDMA scheme when r is large enough to forbid any concurrent transmissions. Thus, the network capacity is a concave curve as a function of r . Similarly, the transport capacity of the network is also a concave curve under different exclusive region radii, and the maximum transport capacity is achieved when $r = 5$ m.

Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space, Table 1
Simulation parameters

Signal bandwidth (W)	500 MHz
Transmission power (P)	0.037 mW
Noise power density (N_0)	-117 dBm/MHz
Path loss exponent (α)	4
Reference distance ($d_{\min}$)	1 m
Path loss at $d_{\min}$ (PL_0)	43.9 dB
Cross correlation (C_0)	0.1-1
Nakagami factor (m)	1~6
Transceiver Efficiency (η)	0.189

Cross-References

- [Area Spectral Efficiency of Ultradense Networks](#)
- [Interference-Aware Scheduling in Wireless Networks](#)
- [Non-Orthogonal Multiple Access \(NOMA\)](#)
- [Random Distance Distribution](#)

References

- Alireza KH, Vinay R, Rudolf R (2006) Broadcast capacity in multihop wireless networks. In: *Proceeding of MobiCom'06*, pp 239–250
- Cai LX, Cai L, Shen X, Mark JW (2009) Capacity analysis of UWB Networks in three-dimensional space. *IEEE/KICS J Commun Netw* 11(3): 287–296
- Cai LX, Cai L, Shen X, Mark JW (2010) REX: a randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna. *IEEE Trans Wirel Commun* 9(1):113–121
- Gamal H (2003) On the scaling laws of dense wireless sensor networks: the data gathering channel. *IEEE Trans Inf Theory* 51(3):1229–1234
- Gastpar M, Vetterli M (2002) The capacity of wireless networks: the relay case. In: *Proceeding of IEEE Infocom'02*
- Grossglauser M, Tse DNC (2001) Mobility increases the capacity of ad-hoc wireless networks. In: *Proceeding of IEEE Infocom'01*, pp 1360–1369
- Gupta P, Kumar P (2000) The capacity of wireless networks. *IEEE Trans Inf Theory* IT-46: 388–404
- Gupta P, Kumar P (2001) Internets in the sky: the capacity of three dimensional wireless networks. *Commun Inf Syst* 1:39–49
- Jafar SA (2005) Too much mobility limits the capacity of wireless ad hoc networks. *IEEE Trans Inf Theory* 51:3954–3965
- Liu K-H, Cai L, Shen X (2008) Exclusive-region based scheduling algorithms for UWB WPAN. *IEEE Trans Wirel Commun* 7(1):933–942
- Radunovic B, Le Boudec J (2004) Optimal power control, scheduling, and routing in UWB networks. *IEEE J Sel Areas Commun* 22(7):1252–1270
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423

Capacity of Wireless Ad Hoc Networks

Pan Li

Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

Synonyms

[Ad hoc networks](#); [Mobile networks](#); [Network capacity](#)

Definitions

Capacity analysis of wireless networks unveils asymptotic performance limits of the networks and can provide guidance to network planning, protocol design, system improvement and upgrade, etc. This chapter introduces the research works on the capacity of wireless ad hoc networks.

Introduction

Network capacity investigation has been intensive in the past couple of decades. A big chunk of work exploring the capacity of wireless networks has appeared in the literature. There are two obvious reasons for the intense attention on this topic. First, network capacity unveils the asymptotic property of network performance. In face of the emerging large-scale networks of connected objects, such as the Internet of Things (IoT), smart city, smart health, and cyber-physical systems, asymptotic capacity is not a cliché but a very critical issue. Second, network capacity predicts network performance limits as a function of the number of nodes in the network, regardless of detailed protocol design. In contrast, as an alternative way to evaluate network perfor-

mance, simulation or numerical results can only be obtained for a certain number of nodes and are hence deterministic. Besides, these results can only be made available after we design all the network protocols considering every detail and may also require a lot of computing resources and time for large-scale networks. Therefore, capacity investigation can provide guidance to network planning, protocol design, system improvement and upgrade, etc., and is interesting, important, and yet very challenging in wireless networks.

Capacity of Static Ad Hoc Networks

In wireless networks, users communicate with each other over a common wireless channel. Due to the wireless broadcast nature, the capacity of wireless networks is constrained by the bandwidth and the node densities. Gupta and Kumar (2000) show that when using omnidirectional antennas, a static random network with n nodes can provide a per-node throughput at $\Theta(W/n \log n)$ bits per second, where W is the channel capacity and $\Theta(x)$ indicates a function on the same order of x . They also show that even under optimal circumstances, the transport capacity is only $\Theta(W/\sqrt{n})$ bit-meters per second for each node. In other words, wireless networks cannot scale when using omnidirectional antennas. Inspired by this seminal paper, many researchers have investigated the capacity issue for wireless networks under various constraints. In the following, we discuss such research works for static homogeneous ad hoc networks and static heterogeneous ad hoc networks, respectively.

Static Homogeneous Ad Hoc Networks

In static homogeneous ad hoc networks, nodes have the same capabilities such as transmission power, communication bandwidth, antennas, traffic demands, and distributions. Li et al. (2001) examine the capacity of wireless ad hoc networks by using simulations based on IEEE 802.11 pro-

ocols. Agarwal and Kumar (2004b) revisit the capacity problem and derive improved capacity bounds for wireless networks. All these works still show that wireless networks using omnidirectional antennas cannot scale. Later, Ozgur et al. (2006, 2007) study the throughput capacity of wireless ad hoc networks and reveal that by intelligent node cooperation and distributed MIMO communication, wireless networks can scale linearly with the number of nodes.

Directional antennas have emerged as a promising technology due to higher spatial reuse ratio, improved communication distance, and reduced interference. Li et al. (2007, 2009a) study the connectivity problem in wireless networks using directional antennas and demonstrate that directional antennas can improve network connectivity. Using the max-flow/min-cut theorem in flow networks, Peraki and Servetto (2003) show that homogeneous random networks with directional antennas can achieve an increase of $\Theta(\log^2 n)$ in maximum stable throughput compared to random networks with omnidirectional antennas (OTOR networks). As the authors point out in the paper, the problem considered there has certain restrictions to the general one considered in Gupta and Kumar (2000). Yi et al. (2003, 2007) also study the capacity improvement of ad hoc networks using directional antennas. Denote by α and β the beamwidth of a transmitter and of a receiver, respectively. For arbitrary networks, by using a simple directional antenna model without the side lobe gain, they show that the throughput capacity gain is $\sqrt{2\pi/\alpha}$ when using directional transmission and omni reception (DTOR), $\sqrt{2\pi/\beta}$ when using omni transmission and directional reception (OTDR), and $2\pi/\sqrt{\alpha\beta}$ when using both directional transmission and directional reception (DTDR), respectively. For random networks, they show that, by ignoring the side lobe gain, the capacity gain can be $2\pi/\alpha$, $2\pi/\beta$, and $4\pi^2/\alpha\beta$, for DTOR, OTDR, and DTDR networks, respectively. Later on, Li et al. (2011) further investigate the capacity

of ad hoc networks using directional antennas under a practical model. Specifically, they prove that the throughput capacity gain of random directional networks using multihop relay schemes compared to random OTOR networks is at most $O(\log n)$ when the side lobe gain of directional antennas is very small, i.e., when $G_s/G_m = o(1)$, G_m and G_s being the main lobe gain and the side lobe gain, respectively. On the other hand, if random directional networks employ one-hop relay schemes, the per-node throughput capacity can be $\Theta(1)$ when the side lobe gain of directional antennas is very small and the number of beams is very large. They also show that the transmission power of each node can be upper bounded by a constant. In addition, the authors present a lower bound on the transport capacity in arbitrary directional networks and find that without side lobe directional antenna gain, the per-node transport capacity scales as $\Omega(N/\sqrt{n})$, $\Omega(\sqrt{N/n})$, and $\Omega(\sqrt{N/n})$, in arbitrary DTDR, OTDR, and DTOR networks, where N is the number of beams.

Static Heterogeneous Ad Hoc Networks

In static heterogeneous ad hoc networks, nodes have different capabilities and/or follow nonhomogeneous distributions. Kozat and Tassiulas (2003), Liu et al. (2003), Agarwal and Kumar (2004a), Zemlianov and Veciana (2005), Pei et al. (2007), Li et al. (2009b), and Zhang et al. (2010) explore the capacity of ad hoc networks with infrastructure support, where there are n uniformly or regularly distributed nodes and m base stations. Their results generally show that the per-node throughput capacity can scale as $\Theta(1)$ if the number of base stations $m = \Theta(n)$. In other words, such a “hybrid wireless network” can only achieve non-diminishing per-node throughput at the cost of very high infrastructure investment. Later on, Li and Fang (2009) investigate the throughput capacity of hybrid wireless networks with various network topologies and traffic patterns. Particularly, they consider n randomly distributed nodes, out of which there are n source

nodes and n^d ($0 < d < 1$) randomly chosen destination nodes, together with n^b ($0 < b < 1$) base stations in a network area of $[0, n^w] \times [0, n^{1-w}]$ ($0 < w \leq 1/2$). They find that only under certain conditions, i.e., when $d > b$ and $d > w$, the maximum achievable throughput is determined by both the number of base stations and the shape of the network area. In all the other cases, the maximum achievable throughput is only constrained by the number of destination nodes. In addition, Li and Fang (2012) consider a heterogeneous wireless network where there are normal nodes and more powerful helping nodes. They show that network capacity is determined by the shape of the network area, the number of destination nodes, the number of helping nodes, and the bandwidth of helping nodes. They also observe that heterogeneous wireless networks can provide throughput higher in the order sense than traditional homogeneous wireless networks only under certain conditions.

All the above works commonly assume that the nodes are uniformly or regularly distributed, while a few works have studied the network capacity when nodes follow nonhomogeneous distributions. Particularly, Alfano et al. (2009) consider that nodes are placed according to a shot-noise Cox process (SNCP), which is essentially a cluster-based power-law distribution model. Li et al. (2012b) assume that nodes are distributed according to a general nonhomogeneous Poisson process (NPP) in a three-dimensional space. Besides, Li et al. (2014) study the multicast capacity for networks with heterogeneous clusters and explore the effect of heterogeneous cluster traffic on the achievable capacity.

Capacity of Mobile Ad Hoc Networks

Realizing that the mobility is an essence of ad hoc networks, Grossglauser and Tse (2002) investigate the capacity of mobile ad hoc networks for the first time and show that a per-node throughput of $\Theta(1)$ can be achieved with a two-hop relaying scheme. However, they have

not addressed the delay-related issues. After that, a big chunk of work explores the capacity and delay in ad hoc networks, considering many different mobility models, such as Brownian mobility model (Lin et al. 2006), random walk model (Gamal et al. 2004, 2006a,b), i.i.d. model (Lin and Shroff 2004; Neely and Modiano 2005; Toupis and Goldsmith 2004), and hybrid model (Sharma et al. 2007). They all show that a constant per-node throughput can be achieved with end-to-end delay bounded by $\Theta(n)$ with possible logarithmic factors. In general, these results demonstrate that we can only either have low throughput and low delay in static networks or have high throughput and high delay in mobile networks. Note that such works assume global mobility, i.e., nodes can move around the whole network. However, this might not always be the case because in many situations nodes only move within a limited region. Li et al. (2012a) employ a practical restricted random mobility model and present smooth trade-offs between throughput and delay by controlling nodes' mobility. Jia et al. (2017) explore optimal capacity-delay trade-off with correlated mobility models.

Moreover, the aforementioned works study the unicast scenario in which each source node only transmits to one destination node. A few works such as Hu et al. (2009) and Wang et al. (2011) discuss the multicast case. Ying et al. (2008) also propose joint coding-scheduling algorithms to achieve the optimal throughput and delay trade-off.

Conclusions

A substantial body of the literature exists addressing the capacity of wireless networks. As wireless technologies advance, there is no doubt that we will have more and more wireless devices surrounding us, making the scale of wireless networks larger and larger. The research on the capacity of wireless networks will become even more important and challenging because of the dynamics and uncertainty of wireless nodes' positions, capabilities, etc., and the security and privacy issues in the networks.

Cross-References

- [Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks](#)
- [Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices](#)
- [Energy Efficiency of Cellular Networks](#)
- [Future Wireless Network Architecture and Network Slicing](#)
- [Interference-aware Distributed Resource Allocation](#)
- [Interference-aware Scheduling in Wireless Networks](#)
- [Mobility in Multicast](#)
- [Network Information Theory](#)

References

- Agarwal A, Kumar P (2004a) Capacity bounds for ad hoc and hybrid wireless networks. *ACM SIGCOMM Comput Commun Rev* 34(3):71–81
- Agarwal A, Kumar P (2004b) Improved capacity bounds for wireless networks. *Wirel Commun Mob Comput* 4(1):251–261
- Alfano G, Garetto M, Leonardi E (2009) Capacity scaling of wireless networks with inhomogeneous node density: upper bounds. *IEEE J Sel Areas Commun (JSAC)* 27(7):1147–1157. Special Issue on Stochastic Geometry and Random Graphs for the Analysis and Design of Wireless Networks
- Gamal A, Mammen J, Prabhakar B, Shah D (2004) Throughput-delay trade-off in wireless networks. In: *Proceedings of the IEEE international conference on computer communications (INFOCOM'04)*, Hong Kong
- Gamal A, Mammen J, Prabhakar B, Shah D (2006a) Throughput-delay trade-off in wireless networks – part I: the fluid model. *IEEE Trans Inf Theory* 52(6):2568–2592
- Gamal A, Mammen J, Prabhakar B, Shah D (2006b) Throughput-delay trade-off in wireless networks – part II: constant-size packets. *IEEE Trans Inf Theory* 52(11):5111–5116
- Grossglauser M, Tse D (2002) Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Trans Netw* 10(4):477–486
- Gupta P, Kumar P (2000) The capacity of wireless networks. *IEEE Trans Inf Theory* 46(2):388–404
- Hu C, Wang X, Wu F (2009) Motioncast: on the capacity and delay tradeoffs. In: *Proceeding of ACM MobiHoc*, New Orleans
- Jia R, Yang F, Yao S, Tian X, Wang X, Zhang W, Xu J (2017) Optimal capacity–delay tradeoff in MANETs

- with correlation of node mobility. *IEEE Trans Veh Technol* 66(2):1772–1785
- Kozat U, Tassiulas L (2003) Throughput capacity of random ad hoc networks with infrastructure support. In: *Proceedings of ACM MobiCom*, San Diego
- Li P, Fang Y (2009) Impacts of topology and traffic pattern on capacity of hybrid wireless networks. *IEEE Trans Mobile Comput* 8(12):1585–1595
- Li P, Fang Y (2012) On the throughput capacity of heterogeneous wireless networks. *IEEE Trans Mobile Comput* 11(12):2073–2086
- Li J, Blake C, Couto DSD, Lee HI, Morris R (2001) Capacity of ad hoc wireless networks. In: *Proceedings of ACM MobiCom*, Rome
- Li P, Zhang C, Fang Y (2007) Asymptotic connectivity in wireless networks using directional antennas. In: *Proceedings of ICDCS*, Toronto
- Li P, Zhang C, Fang Y (2009a) Asymptotic connectivity in wireless ad hoc networks using directional antenna. *IEEE/ACM Trans Netw* 17(4):1106–1117
- Li P, Zhang C, Fang Y (2009b) Capacity and delay of hybrid wireless broadband access networks. *IEEE J Sel Areas Commun (JSAC)* 27(2):117–125. Special Issue on Broadband Access Networks
- Li P, Zhang C, Fang Y (2011) The capacity of wireless ad hoc networks using directional antennas. *IEEE Trans Mobile Comput* 10(10):1374–1387
- Li P, Fang Y, Li J, Huang X (2012a) Smooth trade-offs between throughput and delay in mobile ad hoc networks. *IEEE Trans Mobile Comput* 11(3):427–438
- Li P, Pan M, Fang Y (2012b) Capacity bounds of three-dimensional wireless ad hoc networks. *IEEE Trans Network* 20(4):1304–1315
- Li Y, Peng Q, Wang X (2014) Multicast capacity with max-min fairness for heterogeneous networks. *IEEE Trans Netw* 22(2):622–635
- Lin X, Shroff NB (2004) The fundamental capacity-delay tradeoff in large mobile ad hoc networks. In: *Proceedings of the third annual mediterranean ad hoc networking workshop (MedHoc'04)*, Bodrum
- Lin X, Sharma G, Mazumdar R, Shroff N (2006) Degenerate delay-capacity tradeoffs in ad-hoc networks with brownian mobility. *IEEE/ACM Trans Netw* 14: 2777–2784. Special Issue on Networking and Information Theory
- Liu B, Liu Z, Towsley D (2003) On the capacity of hybrid wireless networks. In: *Proceeding of the IEEE international conference on computer communications (INFOCOM'03)*, San Francisco
- Neely M, Modiano E (2005) Capacity and delay tradeoffs for ad-hoc mobile networks. *IEEE Trans Inf Theory* 51(6):1917–1937
- Ozgur A, Leveque O, Tse D (2006) How does the information capacity of ad hoc networks scale? In: *Proceedings of the forty-fourth annual allerton conference on communication, control and computing*, Monticello
- Ozgur A, Leveque O, Tse D (2007) Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *IEEE Trans Inf Theory* 53(10):3549–3572
- Pei Y, Ambekar V, Modestino J, Wang X (2007) On the throughput capacity of hybrid wireless networks using an l-maximum-hop routing strategy. *Springer Wirel Pers Commun Int J* 42(1):41–48
- Peraki C, Servetto S (2003) On the maximum stable throughput problems in random networks with directional antennas. In: *Proceedings of ACM MobiHoc*, Annapolis
- Sharma G, Mazumdar R, Shroff N (2007) Delay and capacity trade-offs in mobile ad hoc networks: a global perspective. *IEEE/ACM Trans Netw* 15(5):981–992
- Toumpis S, Goldsmith A (2004) Large wireless networks under fading, mobility, and delay constraints. In: *Proceeding of the IEEE international conference on computer communications (INFOCOM'04)*, Hong Kong
- Wang X, Huang W, Wang S, Zhang J, Hu C (2011) Delay and capacity tradeoff analysis for motioncast. *IEEE Trans Netw* 19(5):1354–1367
- Yi S, Pei Y, Kalyanaraman S (2003) On the capacity improvement of ad hoc wireless networks using directional antennas. In: *Proceeding of ACM MobiHoc*, Annapolis
- Yi S, Pei Y, Kalyanaraman S, Azimi-Sadjadi B (2007) How is the capacity of ad hoc networks improved with directional antennas? *Wirel Netw* 13(5):635–648
- Ying L, Yang S, Srikant R (2008) Optimal delay-throughput trade-offs in mobile ad hoc networks. *IEEE Trans Inf Theory* 54(9):4119–4143
- Zemlianov A, Veciana G (2005) Capacity of ad hoc wireless networks with infrastructure support. *IEEE J Sel Areas Commun* 23(3):657–667
- Zhang G, Xu Y, Wang X, Guizani M (2010) Capacity of hybrid wireless networks with directional antenna and delay constraint. *IEEE Trans Commun* 58(7): 2097–2106

Cartel

► Collusion

Car-to-Car Mobile Ad-Hoc Network

► Vehicle Ad-Hoc Networks: Services, Security, and Privacy

CBTC

► Handoff Management in Wireless Communication-Based Train Control Systems

Cell-Free Massive MIMO

Hien Quoc Ngo

School of Electronics, Electrical Engineering
and Computer Science, Queen's University
Belfast, Belfast, UK

Synonyms

Cell-less massive MIMO

Definition

Cell-free massive multiple-input multiple-output (MIMO) is a system comprising a very large number of access points (hundreds or thousands of access points) and a relatively smaller number of active users (tens or hundreds of users) distributed over a large area. The access points use directly measured channel characteristics to jointly serve all users in the same time-frequency band. There are no cells or cell boundaries.

Historical Background

Conventional mobile networks are based on cellular structure where a coverage area is divided into cells, and each cell is served by one access point/base station. These networks are called “cellular wireless networks.” We have seen many revolutions related to the cellular wireless network since its first trial was introduced in the 1970s (Young 1979). Intercell interference is one of the inherent limitations of cellular networks. In particular, users at the cell boundaries perform badly due to the strong intercell interference which limits the performance of the whole networks. To reduce this bottleneck, network MIMO (aka coordinated multipoint with joint transmission, distributed MIMO, and distributed antenna systems) was introduced in the 2000s (Shamai and Zaidel 2001). In network MIMO, the base stations cooperate to serve the users. More precisely, the coverage area is divided into

several disjoint clusters; each is jointly served by several base stations. Basically, network MIMO is equivalent to the cellular network with larger cell size, and each cell is served by multiple base stations. There are still concepts of cells and cell boundaries, and hence, the intercell interference persists. In addition, normally, network MIMO requires complicated signal co-processing with high backhaul overhead and deployment costs.

Cell-free massive MIMO was introduced in 2015 as a practical and scalable version of network MIMO (Ngo et al. 2015). The journal version of (Ngo et al. 2015) was published in 2017 (Ngo et al. 2017). The authors of (Ngo et al. 2015, 2017) showed that cell-free massive MIMO can offer uniformly very good service for all users with very simple maximum-ratio processing. Furthermore, it outperforms the conventional small-cell systems. Recently, cell-free massive MIMO has attracted substantial research interest. The performance of cell-free massive MIMO with zero-forcing processing was studied in (Nayebi et al. 2017) under max-min fairness power controls and in (Nguyen et al. 2017) under energy efficiency optimization. The aspects of channel favorable and channel hardening were investigated in (Chen and Björnson 2017). In (Interdonato et al. 2016), the authors proposed to use beamformed downlink training to estimate the effective channel gains at each user. With this scheme, the system performance improves substantially. The studies of cell-free massive MIMO under practical constraints were exploited in (Huang and Burr 2017), (Bashar et al. 2018), and (Zhang et al. 2017). Papers (Huang and Burr 2017; Bashar et al. 2018) focus on the limited backhaul connections, while paper (Zhang et al. 2017) investigates the effect of hardware impairments on the performance of cell-free massive MIMO. It is shown that under practical constraints, cell-free massive still works well, especially when the number of access points grows large. Another approach to implement cell-free massive MIMO, called user-centric approach, was exploited in (Buzzi and D'Andrea 2017) and Ngo et al. (2018). Both papers show that if each user is served by several selected access points, the

backhaul requirement reduces significantly, while the system performance is still good.

How Cell-Free Massive MIMO Works

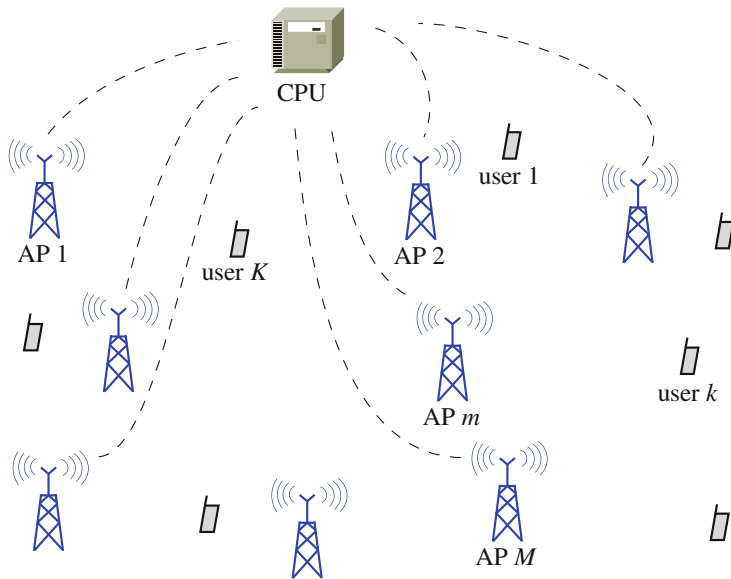
A basic cell-free massive MIMO system is illustrated in Fig. 1. It includes M access points and K users. The access points and users can be equipped with a single or several antennas. Central processing unit (CPU) connects to the access points via a backhaul network to control the data streams and network synchronization and allocate powers.

As in conventional massive MIMO, in cell-free massive MIMO, time-division duplex (TDD) operation is preferable. This is due to the fact that the channel estimation overhead in TDD operation is independent of the number of access points. It depends only on the number of users. As a result, by leveraging TDD, cell-free massive MIMO is scalable in the sense that adding more access points will not affect the operation

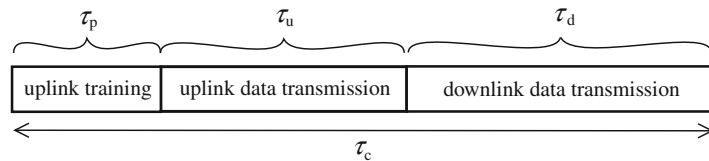
of the channel estimation and, hence, is always beneficial for increased data rate. The TDD transmission protocol is shown in Fig. 2. During a coherence interval of length τ_c symbols, there are three activities: uplink training (τ_p symbols), uplink data transmission (τ_u symbols), and downlink data transmission (τ_d symbols). The uplink channels are estimated at the access points based on the pilot signals transmitted from the users. By virtue of reciprocity, these uplink channel estimates are treated as the downlink channels. Thus, the uplink channel estimates can be used for data detection in the uplink and for processing the transmit signals in the downlink:

- *Uplink training:* in this phase, all users first send their pilot sequences of length τ_p symbols to the access points. Then from the received pilot signals, each access point estimates all its channels to the users. Typically, the linear minimum mean-square error estimation technique is used. If the coherence interval is large (relative to the

Cell-Free Massive MIMO,
Fig. 1 Cell-free massive MIMO system



Cell-Free Massive MIMO,
Fig. 2 TDD transmission protocol



number of users K), then we can choose $\tau_p \geq K$, and hence, the pilot sequences assigned for the K users can be pairwise orthogonal. Otherwise, non-orthogonal pilot sequences have to be used throughout the network. The channel estimate of a given user is contaminated by pilot signals transmitted from other users. This degrades the system performance, even when the number of access points is very large. This effect is called “pilot contamination effect”.

- *Uplink payload data transmission*: in this phase, all users simultaneously send their signals to the access points. Each access point uses its local channel estimates to process the received signals and sends the processed version to the CPU. The CPU then detects all signals transmitted from all K users. Simple linear processing such as maximum-ratio processing (aka matched filtering) is recommended at each access point. Alternatively, signal processing can be done at the CPUs. With this scheme, the APs need to forward their channel estimates and received data to the CPU for signal detection.
- *Downlink payload data transmission*: in this phase, the access points use their local channel estimates to precode the symbols intended for all K users (these symbols are sent from the CPU) and broadcast the precoded versions to the users. Each user then detects the desired symbol from the received signal. Simple linear processing such as maximum-ratio processing (aka conjugate beamforming) is recommended at each access point.

Why Cell-Free Massive MIMO

Cell-free massive MIMO is the combination between distributed structure and massive MIMO technology. Thus, it reaps all benefits obtained from these systems:

- *Uniformly good service for all users in the network*: in cell-free massive MIMO, the access points are spread out over a large area. In other words, many access points are brought close to the users which yields low path losses and

an increased macro-diversity. As a result, it can offer a very high probability of coverage and uniformly good service for all users.

- *Huge spectral and energy efficiency*: by exploiting the favorable propagation property when the number of access points is large (the channel vectors between the users and all access points are (nearly) pairwise orthogonal), many users can be served simultaneously in the same frequency band with small inter-user interference. Thus, a huge spectral efficiency can be obtained. In addition, with massive number of access points, cell-free massive MIMO can focus the energy to the locations of its desired users. As a result, it can offer very high energy efficiency.
- *Simple signal processing*: owing to the favorable propagation, cell-free massive MIMO can achieve very good performance with simple linear processing such as maximum-ratio or zero-forcing processing. In addition, from the law of large numbers, uncorrelated noise and small-scale fading can be averaged out. Therefore, the system design (e.g., power controls and pilot assignments) depends only on the large-scale fading.

Example

To clarify the benefits of cell-free massive MIMO mentioned above, consider the downlink transmission of a simple cell-free massive MIMO system where M access points serve two users. All access points and users are equipped with a single antenna. It is assumed that the access points have perfect channel state information (the uplink training is perfect) and use maximum-ratio processing to precode the signals.

Denote by g_{mk} the channel between the m th access point and the k th user. The channel g_{mk} models large-scale fading (path loss and shadowing) β_{mk} and small-scale fading h_{mk} as follows:

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}. \quad (1)$$

We assume that h_{mk} is Gaussian distributed with zero mean and unit variance. Let q_k , where

$\mathbb{E}\{|q_k|^2\} = 1$, be the symbol intended for the k th user. Then, with maximum-ratio processing, transmitted signal from the m th access point is

$$x_m = \sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}} (g_{m1}^* q_1 + g_{m2}^* q_2). \quad (2)$$

The factor $\sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}}$ guarantees that the normalized transmit power of the access point is ρ , i.e., $\mathbb{E}\{|x_m|^2\} = \rho$. To precode the symbols q_1 and q_2 as in (2), the m th access point only needs to know its local channels g_{m1} and g_{m2} which can be estimated during the training phase. Thus, channel state information does not need to be shared among the access points.

Focus on user 1. The same analysis can be done for user 2. Since all M access points simultaneously broadcast their signals, user 1 receives

$$y_1 = \sum_{m=1}^M g_{m1} x_m + n_1, \quad (3)$$

where n_1 is the additive noise at user 1. We assume that n_1 is Gaussian distributed with zero mean and unit variance. Substituting (2) into (3), we obtain:

$$\begin{aligned} y_1 &= \sum_{m=1}^M g_{m1} \sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}} \\ &\quad (g_{m1}^* q_1 + g_{m2}^* q_2) + n_1 \\ &= \underbrace{\sum_{m=1}^M \sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}} |g_{m1}|^2 q_1}_{\text{desired signal}} \\ &\quad + \underbrace{\sum_{m=1}^M \sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}} g_{m1} g_{m2}^* q_2}_{\text{interference and noise}} + n_1. \end{aligned} \quad (4)$$

By Chebyshev's theorem (which is the general version of the law of large numbers), when the number of access points goes to infinity, we have:

$$\frac{\text{desired signal}}{M}$$

$$- \frac{1}{M} \sum_{m=1}^M \sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}} \beta_{m1} q_1 \xrightarrow{P} 0, \quad (6)$$

$$\frac{\text{interference and noise}}{M} \xrightarrow{P} 0, \quad (7)$$

where $\xrightarrow{P}$ denotes convergence in probability.

It can be seen from (6) and (7) that, when the number of access points grows large, the desired signals are added up coherently from all access points, while the interference is averaged out. As a result, the received signal at user 1 normalized by M is

$$\frac{y_1}{M} - \frac{1}{M} \sum_{m=1}^M \sqrt{\frac{\rho}{\beta_{m1} + \beta_{m2}}} \beta_{m1} q_1 \xrightarrow{P} 0. \quad (8)$$

This implies that when M is large, the interference from user 2 and noise are canceled out. The desired signal normalized by M includes only the desired symbol q_1 . Unlimited capacity can be achieved.

Challenges and Open Problems

Despite the huge advantages of cell-free massive MIMO, many challenges still need to be tackled:

- **Backhaul requirements:** cell-free massive MIMO requires huge backhaul connections and signaling, especially when the number of access points is very large. Therefore, suitable transmission approaches should be investigated to make cell-free massive MIMO feasible for practical implementations. One possible approach is the user-centric approach. With user-centric approach, each user is served by some chosen access points around it. In addition, multiple CPUs can be used.
- **Synchronization:** the system needs to be synchronized so that all access points can coherently serve all users. In cell-free massive MIMO, since the numbers of access points and users are very large, and the access points/users are located randomly in a very large area, the network synchronization is

very challenging. New simple methods which can synchronize the cell-free massive MIMO networks with low cost should be investigated.

- **Channel state information acquisition:** acquisition of channel state information is very important in cell-free massive MIMO. This is done via the uplink training under TDD operation. In general, the number of users is large, while the coherence interval is limited. Thus, it does not guarantee that the pilot sequences assigned for all users are pairwise orthogonal. This causes pilot contamination effect and degrades the system performance significantly. Therefore, good pilot assignment methods or/and good channel estimation algorithms to reduce the pilot contamination effect need to be exploited.

Key Applications

Cell-free massive MIMO can be applied in every wireless networks, particularly in mobile networks, device-to-device communications, and machine-type communications such as the Internet of Everything, Internet of Things, and Smart X.

Cross-References

- [Coordinated Multipoint Transmission and Reception](#)
- [Massive MIMO](#)
- [Massive MIMO Channel Estimation](#)
- [Network Slicing-Enabled Green C-RAN](#)
- [Network MIMO](#)

References

- Bashar M, Cumanan K, Burr AG, Debbah M, Ngo HQ (2018) Cell-free massive MIMO with limited backhaul. In: Proceedings of IEEE international conference on communications (ICC), Beijing
- Buzzi S, D'Andrea C (2017) Cell-free massive MIMO: User-centric approach. *IEEE Wireless Commun Lett* 6(6):706–709. <https://doi.org/10.1109/LWC.2017.2734893>

- Chen Z, Björnson E (2017) Channel hardening and favorable propagation in cell-free Massive MIMO with stochastic geometry. CoRR abs/1710.00395. <http://arxiv.org/abs/1710.00395>
- Huang Q, Burr A (2017) Compute-and-forward in cell-free massive MIMO: Great Performance with Low backhaul load. In: Proceedings of IEEE international conference on communications (ICC), Paris
- Interdonato G, Ngo HQ, Larsson EG, Frenger P (2016) How much do downlink pilots improve cell-free massive MIMO? In: IEEE global communications conference (GLOBECOM), pp 1–7. <https://doi.org/10.1109/GLOCOM.2016.7841875>
- Nayebi E, Ashikhmin A, Marzetta TL, Yang H, Rao BD (2017) Precoding and power optimization in cell-free massive MIMO systems. *IEEE Trans Wireless Commun* 16(7):4445–4459
- Ngo HQ, Ashikhmin AE, Yang H, Larsson EG, Marzetta TL (2015) Cell-free massive MIMO: Uniformly great service for everyone. In: Proceedings of IEEE SPAWC, Stockholm, Sweden
- Ngo HQ, Ashikhmin A, Yang H, Larsson EG, Marzetta TL (2017) Cell-free massive MIMO versus small cells. *IEEE Trans Wireless Commun* 16(3):1834–1850
- Ngo HQ, Tran LN, Duong TQ, Matthaiou M, Larsson EG (2018) On the total energy efficiency of cell-free massive MIMO. *IEEE Trans Green Commun Netw* 1(2):25–39
- Nguyen LD, Duong TQ, Ngo HQ, Tourki K (2017) Energy efficiency in cell-free massive MIMO with zero-forcing precoding design. *IEEE Commun Lett* 8(21):1871–1874
- Shamai S, Zaidel BM (2001) Enhancing the cellular downlink capacity via co-processing at the transmitting end. In: Proceedings of IEEE VTC-Spring, vol 3, Atlantic City, pp 1745–1749
- Young WR (1979) Advanced mobile phone service: introduction, background, and objectives. *Bell Syst Tech J* 58(1):1–14
- Zhang J, Wei Y, Björnson E, Han Y, Li X (2017) Spectral and energy efficiency of cell-free massive MIMO systems with hardware impairments. In: Proceedings of international conference on wireless communications and signal processing (WCSP), Nanjing

Cell-Less Massive MIMO

- [Cell-Free Massive MIMO](#)

Cellular and Mobile Communication Networks

- [Cellular Networks: An Evolution from 1G to 4G](#)

Cellular Networks: An Evolution from 1G to 4G

Indika A. M. Balapuwaduge and Frank Y. Li
University of Agder, Kristiansand, Norway

Synonyms

Cellular and mobile communication networks

Definitions

A cellular network is a wireless communication network which provides services to mobile users covered by a region consisting of multiple cells. Frequency reuse is a fundamental concept in cellular networks.

Historical Background

Prior to the invention of cellular technologies, there were few mobile telephone systems in the late 1940s in the USA and in the early 1950s in Europe, such as car-based telephone systems. The push-to-talk technique was used in these systems, and communication was performed over a single channel. With a single channel and a single antenna at each device, the transmission and reception of signals were performed in a half-duplex manner. Due to low capacity, restricted mobility support, and poor service quality, those systems were not efficient. Later on, mobile telephone system (MTS) and improved MTS (IMTS) were introduced in order to support a larger number of mobile stations. IMTS adopted two channels, one for sending and one for receiving, allowing mobile communication to work in a full-duplex mode. However, the concept of *cells* did not apply to these pre-1G mobile communication systems.

The breakthrough for increasing capacity appeared in the 1970s when the technique for *frequency reuse* was developed by dividing a coverage region into multiple cells and reusing

the same frequency at different cells which are sufficiently far away from each other. In 1979, the first 1G cellular network was launched in Japan. In 1981, the first international cellular network, Nordic Mobile Telephone (NMT) systems, came to operation in Nordic countries. In 1983, two other 1G systems, Advanced Mobile Phone System (AMPS) and Total Access Communication System (TACS), were introduced in the US and other European countries including the UK and Italy, respectively.

Through a journey which started about 40 years ago, cellular communications have experienced dramatic changes and upgrades from 1G to 4G. Those transitions include the system change from analog to digital, the core network evolution from circuit-switched to packet-switched technologies, and its further transfer toward all Internet Protocol (IP) networks.

First-Generation Cellular Systems

The 1G mobile systems were designed for providing voice services only and were developed based on analog technologies. For medium access of multiple users, frequency-division multiple access (FDMA) schemes were adopted. For carrying voice traffic, frequency modulation (FM) was employed as the modulation scheme.

Nordic Mobile Telephone Systems

NMT is a mobile telephone network covering first the five Nordic countries, Denmark, Finland, Norway, Sweden, and Iceland, and later quite a few other countries. NMT has two variants based on the operational frequency bands, known as NMT-450 and NMT-900, respectively. NMT-450 was developed in order to establish a compatible telephone system in the Nordic countries (Nordic Mobile Telephone Group 1995). Initially NMT-450 targeted at deploying macro cells at the 450 MHz band to provide larger cell coverage, and later it was modified to operate in the 900 MHz band by considering the size and transmission power constraint of handsets. NMT systems were initially launched in Norway and

Sweden as a national service, and later on it was enhanced with roaming services across countries. NMT-900 bears more channels than the NMT-450 network, able to serve a higher number of subscribers.

Advanced Mobile Phone System

The AMPS systems were deployed in North America. It has 2×20 MHz bandwidth within the 800–900 MHz frequency band (825–845 MHz for uplink and 870–890 MHz for downlink traffic respectively) allocated by the Federal Communications Commission (FCC). With 30 kHz bandwidth for each channel, it provides 832 duplex channels. This system employs seven-cell clusters for frequency reuse and uses mainly 120° sector antennas. Three sectors for a single AMPS cell site were designed to achieve a carrier-to-interference ratio of 18 dB with satisfactory voice quality. In 1991, an improved version of AMPS named as narrowband AMPS (N-AMPS) was developed to further increase AMPS capacity with additional advanced features such as authentication and caller ID.

Total Access Communication Systems

TACS was the first standard that used the 900 MHz band and was deployed in other European countries (Goldsmith 2005). TACS was operated at a higher frequency than AMPS and adopted narrower bandwidth per channel (25 kHz in TACS versus 30 kHz in AMPS). With narrower bandwidth for each channel, the total number of channels is increased for a band with fixed bandwidth. Operated at a higher frequency, the coverage for each cell will be reduced given the same transmission power and channel condition. In other words, this system was designed for having higher capacity rather than coverage by deploying a larger number of cells and allowing lower transmission power for mobile stations. Indeed, TACS was proved to be efficient and economical for highly dense urban areas. A variant of TACS, J-TACS, was also adopted in Japan.

Second-Generation Cellular Systems

The core network of both 1G and 2G cellular networks was built based on circuit-switched technologies, and the service was mainly targeted at voice traffic. However, unlike 1G systems, 2G employed *all-digital* transmission technologies for both control signaling and data traffic. The introduction of digital communication provides a series of important features such as the support of advanced source and channel coding, more efficient spectrum utilization, and a high degree of resistance against interference and channel fading. In addition, the handling of control information is more efficient in digital systems.

2G cellular networks were deployed worldwide. They are represented by four major standards, i.e., Global System for Mobile communications (GSM), Interim Standard (IS)-136 or Digital AMPS (D-AMPS), IS-95 or cdmaOne, and Personal Digital Cellular (PDC). Except cdmaOne, the other three systems are based on time-division multiple access (TDMA) mechanisms for medium access. TDMA allows multiple users to share the same channel in the frequency domain by allocating a specific short period of time, known as time slot, to each user for their channel access. Among these four standards, GSM is the most popular 2G technology. Given the fact that 3G and 4G are already widely deployed in the world as of 2017, GSM still occupies approximately 39% of the global mobile market share.

Global System for Mobile Communications

In 1982, a committee called *Groupe Spécial Mobile* was established in order to design a Pan-European digital cellular standard for mobile communications which would replace the incompatible 1G analog systems. The European Telecommunications Standards Institute (ETSI) initiated the development of the first version of GSM, and the first GSM call was made in 1991. The success of GSM is represented by its over 90% market share in the 2G world, covering globally around 200 countries and territories.

GSM is operated mainly in three frequency bands, i.e., GSM-900, GSM-1800, and GSM-

1900 with 124, 374, and 299 radio channels, respectively. The original GSM network was developed for voice communication supporting a variety of voice codecs. The bandwidth for each GSM channel is 200 kHz, and there are eight time slots in each frame which lasts for 4.615 ms. The raw data transmission rate achieved at each channel is 270.833 kbps, and it is shared by eight users. With the combination of TDMA and FDMA, GSM provides the capability of simultaneous conversations at the same frequency through different time slots. Other salient features of GSM include data encryption, subscriber identity module (SIM) card which gives a unique identity to each mobile station, and global roaming (Eberspächer et al. 2009). Furthermore, GSM is enhanced to provide short message service (SMS), however, still based on circuit-switched technologies.

GPRS and EDGE

To support higher data transmission rates in GSM networks and provide IP services without replacing the network infrastructure, developments were made to upgrade GSM networks into General Packet Radio Service (GPRS) in year 2000. GPRS is also known as 2.5G, and there are two major enhancements for the evolution from GSM to GPRS. At the radio access network, a user is allowed to occupy up to five time slots so that 114 and 20 kbps are achieved for downlink and uplink traffic, respectively. At the core network, two entities, Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN), are introduced. With the support of Packet Data Protocol (PDP) and GPRS Tunneling Protocol (GTP), end-to-end IP services can be provided.

In 2003, there was another improvement in GSM systems with the deployment of Enhanced Data rates for GSM Evolution (EDGE) networks. EDGE is regarded as an extension of the GPRS radio access network through advanced modulation schemes. It increased the data rates to 384 kbps for downlink and 60 kbps for uplink, respectively.

Third-Generation Cellular Systems

While GSM was in its early stage, the standardization of the next-generation mobile telecommunication network was also initiated by ETSI aiming to develop a new system called Universal Mobile Telecommunications System (UMTS). High spectrum efficiency, data rates up to 2 Mbps, variable bit rates, QoS requirements based on service types, support of asymmetric uplink and downlink traffic, and coexistence with 2G systems are the main requirements for 3G systems (Holm and Toskala 2007). Meanwhile ITU commenced proposing recommendations for 3G systems known as International Mobile Telecommunications 2000 (IMT-2000) and started to investigate suitable spectrum for 3G. The establishment of the 3rd Generation Partnership Project (3GPP) in 1998 became a milestone for the standardization of cellular networks as a *global* standard from 3G and beyond. In Table 1, we summarize 3GPP releases and their main features. Note that current 4G and upcoming 5G standards are also standardized by 3GPP.

The key mechanism used for medium access in 3G is CDMA which allows multiple mobile stations to transmit at the same time and in the same frequency band. In a CDMA cell, each subscriber is assigned a unique code, and the codes assigned to different stations are orthogonal to each other. As such, the number of simultaneous calls in a CDMA cell is soft limited (based on the interference level), not hard limited as in GSM.

UMTS WCDMA

Wideband CDMA (WCDMA) was the air interface for the UMTS standard originally proposed by ETSI in 1998. To provide peak data rates from 384 kbps to 2.048 Mbps, the WCDMA system operates on wider channels each with 5 MHz bandwidth. However, the core network architecture of UMTS remains the same as the existing GSM/GPRS networks.

WCDMA is a direct-sequence spread spectrum (DSSS) system which initially operates in the 1885–2025 MHz and 2110–2200 MHz frequency bands for uplink and

Cellular Networks: An Evolution from 1G to 4G, Table 1 3GPP release dates and features

3GPP release	Start/release date	Summary of key features
R99	1996/2000	First release of the UMTS standard
R4	1998/2001	This release added features including an all-IP core network. It was originally referred to as Release 2000
R5	2000/2002	IP Multimedia Subsystem and High-Speed Downlink Packet Access (HSDPA)
R6	2000/2004	Integrate the operation of UMTS with wireless LAN networks and added enhancements to IMS (including push-to-talk over cellular), and it added High-Speed Uplink Packet Access (HSUPA)
R7	2003/2007	Detailed improvements to QoS for applications such as VoIP and upgrades for High-Speed Packet Access (HSPA+) Evolution as well as changes for EDGE
R8	2006/2008	Provide details for the LTE System Architecture Evolution, SAE, an all-IP flat network architecture providing the capacity and low latency required for Long-Term Evolution (LTE) and future evolutions
R9	2008/2009	Further enhancements to the SAE as well as allowing for WiMAX and LTE/UMTS interoperability
R10	2009/2011	Up to 3 Gbps downlink and 1.5 Gbps uplink, carrier aggregation (CA), relay nodes to support heterogeneous networks, higher-order MIMO antenna configurations
R11	2010/2012	Enhancements to carrier aggregation, MIMO, relay nodes, coordinated multi-point transmission and reception to enable simultaneous communication with multiple cells, introduction of new frequency bands
R12	2011/2015	Enhanced small cells for LTE, inter-site carrier aggregation, interworking between LTE and Wi-Fi or HSDPA
R13	2012/2016	LTE-U, LTE for machine-type communication (MTC), full-dimension MIMO, LTE Advanced Pro
R14	2015/2017	Energy efficiency, location services, mission-critical data and video, massive IoT
R15	2016/2018	5G Phase 1 (new radio)
R16	2017/2019-2020	5G Phase 2

downlink, respectively. It supports operation modes of both frequency-division duplex (FDD) when symmetric uplink/downlink channels are available and time division duplex (TDD) when only asymmetric spectrum is available. In addition to achieving higher data rate, the support for multi-code operation, larger number of spreading factors, and enhanced transmission diversity are the key factors which lead WCDMA to the most popular 3G technology. Moreover, for the purpose of leveraging the GSM coverage for WCDMA, seamless handovers between GSM and WCDMA are also supported as well as dual-mode handsets.

High-Speed Packet Access (HSPA)

HSPA includes two phases as the beyond UMTS WCDMA enhancements by 3GPP, i.e., HSDPA

in R5 and HSUPA in R6. The main idea of HSDPA is to increase downlink data rate through techniques including adaptive modulation and coding (AMC), hybrid automatic repeat request (HARQ), and fast packet scheduling. Unlike in the previous standards, the medium access control (MAC) layer of HSDPA systems is installed at the base station, i.e., NodeB. Thus, retransmissions can be controlled directly by NodeB, leading to faster retransmission and accordingly shorter delay with packet data operation (Holm and Toskala 2007). HSDPA is capable of supporting up to 14.4 Mbps peak theoretical throughput. Similarly HSUPA is developed to support enhanced packet data throughput for uplink. HSUPA enables additional features such as fast power control and variable spreading factors which are disabled in HSDPA.

However, HSUPA does not support AMC. It is capable of providing up to 5 Mbps peak uplink throughput.

Fourth-Generation Cellular Systems

The data rate provided by 3G networks was not able to meet the growing demand for Internet access via mobile phones. From 2008, ITU's Radiocommunications Sector (ITU-R) initiated the process for developing a new system known as IMT-Advanced. The main requirements for this new system include supporting 100 Mbps and 1 Gbps peak data rate for high- and low-mobility scenarios, respectively, bandwidth scalability up to 100 MHz, mobility support with up to 350 km/h, 10 ms user plane latency and 100 ms control plane latency, worldwide roaming capability, inter-networking with other 2G and 3G systems, and improved spectral efficiency (Korhonen 2014). Radio technologies that could meet these requirements are termed as 4G systems. LTE Advanced (LTE-A) developed by 3GPP is regarded as the de facto 4G mobile communications system from a global perspective. Although LTE does not meet the requirements for 4G as specified by ITU, it is often marketed as a 4G technology.

LTE

LTE is presented here under the 4G umbrella considering the fact that both LTE and, the *true* 4G technology, LTE-A, are based on orthogonal frequency-division multiplexing (OFDM)/OFDM access (OFDMA) mechanisms for medium access. LTE supports both FDD and TDD operations and provides flexible operations in both symmetric and asymmetric spectra. To enhance uplink power efficiency, LTE adopts single-carrier frequency-division multiple access (SC-FDMA). To achieve higher data rate, LTE offers flexible bandwidth allocation up to 20 MHz. Moreover, much shorter frame sizes (10 ms frames and 1 ms sub-frames) are introduced. With a configuration of 20 MHz spectrum and 4×4 MIMO, 326 Mbps on the

downlink and 86 Mbps on the uplink can be achieved in LTE radio networks.

To upgrade LTE core networks, 3GPP R8 introduced Evolved Packet Core (EPC) which was designed to provide higher capacity, all-IP support, and reduced latency. Unlike the hierarchical architecture used in GPRS and UMTS core networks, the main design principle in EPC was to keep the architecture simple and flat.

LTE Advanced

The LTE-A standard ratified by 3GPP R10 is the major standard for 4G. Many of the existing features in R8 are inherently supported in LTE-A. In addition, carrier aggregation up to five component carriers, use of relays, higher-order MIMO, and enhanced inter-cell interference coordination (eICIC) served as new features in LTE-A. Furthermore, in R11, coordinated multipoint (CoMP) transmission and reception and enhanced self-organizing network (SON) are designed as parts of the LTE-A networks.

The same as in LTE, the medium access mechanism in LTE-A adopts OFDMA for downlink and SC-FDMA for uplink. The main advantages of OFDMA include the ability to gain much frequency diversity through randomly distributed subcarriers, multiuser diversity via assigning contiguous sets of subcarriers, and adaptive bandwidth allocation. To increase transmission power efficiency and reduce the cost of power amplifiers for mobile stations, SC-FDMA is employed thanks to its low peak-to-average power ratio. At the physical layer, adaptive coding and modulation (ACM) is adopted in 4G systems. With ACM, the modulation order and the coding rate are fine-tuned based on the channel state information to gain the full use of radio channels.

Another technique for data rate enhancement is channel aggregation which assembles multiple channels together to perform data transmission over wider bandwidth. Furthermore, LTE-A continues to improve capacity and reliability through more advanced MIMO technologies. While single-site MIMO introduces beamforming, spatial multiplexing, and transmit diversity into the system, cooperative MIMO facilitates CoMP transmission and reception. Under CoMP,

a mobile station is able to receive signals from multiple base stations and likewise for the uplink transmission. By adopting a proper coordination scheme, the received signal quality is greatly improved.

Due to the fact that mobile stations cannot support MIMO with many antennas, cell-edge users may not obtain 4G-level quality of service, although a large number of antennas can be installed at the base station. To improve the performance of these users, relaying is adopted as a cooperative communication technique where single-antenna mobile stations transmit their signals to the base station via a relay station that is located much closer to the cell edge (Sesia et al. 2011). Another critical requirement for IMT-Advanced is effective power management. The aforementioned techniques including SC-FDMA, advanced multi-antenna design, and relaying collectively contribute to significant reduction of mobile station power consumption.

Key Applications

On its journey from 1G to 4G, cellular networks experienced FDMA-based analog systems (NMT, AMPS, TACS), TDMA-based digital systems (GSM, GPRS), CDMA-based IP-enabled systems (WCDMA, HSPA), and OFDM-based broadband mobile systems (LTE/LTE-A). The core network of cellular systems has gradually evolved from a circuit-switched telephone network to a packet-switched mobile network with an all-IP capability. Another trend along with this evolution is that nowadays cellular technologies are being developed as a global common standard in lieu of multiple regional standards developed in the early years which were not compatible with each other.

Cross-References

- [Analog and Digital Communications](#)
- [Key Technologies in 4G/LTE Network](#)
- [Random Access Technology in Cellular Networks](#)

References

- Eberspächer J, Vögel HJ, Bettstetter C, C Hartmann C (2009) *GSM architecture, protocols and services*. Wiley, West Sussex
- Goldsmith A (2005) *Wireless communications*. Cambridge University Press, Cambridge
- Holm H, Toskala A (2007) *WCDMA for UMTS: HSPA evolution and LTE*. Wiley, West Sussex
- Korhonen J (2014) *Introduction to 4G mobile communications*. Artech House, London
- Nordic Mobile Telephone Group (1995) *Nordic mobile telephone*. Stockholm
- Sesia S, Toufik I, Baker M (2011) *LTE – the UMTS long term evolution: from theory to practice*. Wiley, West Sussex

Cellular Technologies

- [Evolution of Vehicular Networks in the Transition from 3G to 4G Systems](#)

Cellular Unlicensed Spectrum Technology

Jeongho Jeon and Tom Cruz
Intel Corporation, Santa Clara, CA, USA

Synonyms

[License-Assisted Access. \(Not Licensed-...\)](#)

Definition

Cellular unlicensed spectrum technology is a collective term that refers to cellular technologies exploiting the unlicensed band spectrum.

Historical Background

The first and foremost example was designed by the 3rd Generation Partnership Project (3GPP),

which is Long-Term Evolution (LTE) License-Assisted Access (LAA) technology (Kwon et al. 2017). LAA received its name from the system's design, in which it only allows the use of unlicensed carriers when it is operated together with a licensed carrier. LTE LAA has been developed in multiple phases: Rel-13 LAA Study Item (SI) phase, Rel-13 LAA Work Item (WI) phase, Rel-14 enhanced LAA (eLAA) WI phase, and Rel-15 further enhanced LAA (FeLAA) WI phase. FeLAA in Rel-15 is currently ongoing. Similar to 3GPP, MulteFire has been specifying the stand-alone unlicensed system, a system designed to operate on unlicensed carriers without the need of a licensed carrier ([Online]. Available: <http://multefire.org>). MulteFire is an additional feature specification to the 3GPP LTE LAA design to enable stand-alone unlicensed operation. Note that MulteFire is not a part of 3GPP. On the other hand, 3GPP is currently in the phase of developing New Radio (NR), the next generation radio access technology for International Mobile Telecommunications (IMT) 2020. NR is being developed assuming the use of licensed carrier in the first phase. It is planned to start the feasibility of operating NR using unlicensed spectrum in late 2017 (RP-170828 2017).

Standardization History

3GPP Rel-13 LAA SI

The objective of the Rel-13 LAA SI was to study the feasibility of LTE operation in unlicensed spectrum with the focus on fair coexistence with incumbent systems, e.g., Wi-Fi, and meeting the regulatory requirements (Technical Report 36.889 2015). An extensive simulation campaign had been conducted with the participation of both cellular and Wi-Fi industries. The evaluation results showed that the LAA technology can indeed fairly coexist with Wi-Fi networks for various traffic types including, but not limited to, File Transfer Protocol (FTP) and Voice over IP (VoIP) traffic, if a listen-before-talk (LBT) mechanism, similar to that of Wi-Fi, is augmented to LAA. Accordingly, the conclusion

of the SI was that "it is feasible for LAA to fairly coexist with Wi-Fi and other LAA networks if an appropriate channel access scheme is adopted." The LAA SI phase took place during the first half of 3GPP Rel-13 from late 2014 to mid-2015.

3GPP Rel-13 LAA WI

As the LAA SI transitioned into a WI, the objective became the specification of LTE Downlink (DL) operation using unlicensed secondary cell(s) (SCell) with a licensed primary cell (PCell) (RP-151045 2015). The LAA technology is a global single solution, which was designed to meet different regional regulatory requirements. Just as in Rel-13 LAA SI, ensuring the fair coexistence with existing Wi-Fi networks remained as the highest priority. Consequently, the specification of the channel access mechanism, including the LBT procedure, was the main focus. The LAA WI was conducted during the second half of 3GPP Rel-13, beginning in mid-2015 and completed in late 2015.

3GPP Rel-14 eLAA WI

Following the Rel-13 LAA WI, a new WI was created with the objective to enable the LTE Uplink (UL) operation on the unlicensed carrier in addition to the DL operation defined in Rel-13. The Rel-14 LAA WI was titled enhanced LAA (eLAA) (RP-152272 2015). The channel access mechanism for the UL fundamentally resembles that of the DL. The LBT mechanism, which was first incorporated in Rel-13 for the DL, was now tailored to operate for UL. In addition, the channel occupancy sharing mechanism between the DL and UL was defined. The UL control channel and the random access channel for the initial access were not defined for an unlicensed carrier as they stay associated with a licensed UL carrier. The eLAA WI was conducted during 3GPP Rel-14 from early 2016 to late 2016.

3GPP Rel-15 FeLAA WI

3GPP is currently conducting the specification for further enhanced LAA (FeLAA) WI within the Rel-15 time frame, and it is scheduled to finish by early 2018 (RP-170848 2017). One objective of the FeLAA WI is to specify the sup-

port for multiple starting and ending positions in a subframe for UL and DL on an unlicensed carrier. This new design would further enhance the coexistence with neighboring systems by improving the efficiency of unlicensed resource utilization by defining finer granular starting and ending positions of a transmission by LAA. Another objective is to study and specify the autonomous uplink access, which does not rely on a UL scheduling grant for transmission. This objective was motivated to enhance the LAA UL performance from that of Rel-14.

Key Technical Components

Listen-Before-Talk

The adoption of LBT in LAA system was crucial in achieving fair coexistence with neighboring systems sharing the unlicensed spectrum in addition to the regulatory requirements. The LBT mechanism by LAA fundamentally resembles the WLAN's carrier sense multiple access with collision avoidance (CSMA/CA) (IEEE Std 802.11TM-2012 [2012](#)). A node that intends to transmit first performs channel sensing before transmission. The purpose of the additional random backoff mechanism is to avoid collisions when more than one node senses the idle channel and transmits at the same time. An important component of LBT design is the choice of an energy detection (ED) threshold, which translates into the level of sensitivity to detect the neighboring ongoing transmissions. The impact of an ED threshold on the fair coexistence can be found in (Jeon et al. [2015](#)).

Discovery Reference Signal

The discovery reference signal (DRS) was introduced in LTE Rel-12 for radio resource measurement (RRM) during the small cell OFF state. The DRS consists of synchronization signals and reference signals. Note that in the licensed carrier, synchronization signals and reference signals are always transmitted regardless of whether there is actual data transmission or not during the small cell ON state; however, in the unlicensed carrier, if there is no data transmission, then there is no

transmission at all. Therefore, the use of DRS on an LAA SCell is critical to maintain a reliable connection. Like any transmission on unlicensed carrier, the transmission of DRS is also subject to LBT.

Interlaced UL Resource Allocation

European Telecommunications Standards Institute (ETSI) specifies that the occupied channel bandwidth (OCB) shall be between 80% and 100% of the declared nominal channel bandwidth (ETSI EN 301 893 [2017](#)). To meet the OCB regulation by ETSI, the eLAA allocates user equipment (UE) with equally spaced resource blocks (RBs) spread across the entire bandwidth for UL transmission. In more detail, each interlaced resource allocation consists of ten equally spaced RBs for both 10 and 20 MHz system bandwidth.

Partial Subframes

In unlicensed spectrum, the LBT procedure can finish at any time subject to the ongoing transmission and/or the random backoff in LBT. To efficiently utilize the radio resources, the concept of a partial subframe has been introduced in Rel-13 LAA, where DL transmission can start at the first or second slot boundaries of a subframe. The DL transmission also may not end at the subframe boundary due to the limited channel occupancy time. To utilize the ending DL partial subframes with minimal specification efforts, the existing TDD special subframe structure is reused. In the current Rel-15 FeLAA WI, further DL/UL partial subframe options are to be added.

Autonomous UL Access

In the Rel-15 FeLAA WI, one of the objectives here is autonomous UL access. It was observed by multiple sources that the UL performance can be impacted when operated in unlicensed spectrum. A possible explanation can be that multiple rounds of contention are imposed before a UE can transmit on the UL. That is, LBT needs to be performed at the base station before sending the UL grant. A particular UE has to be exclusively scheduled for a given resource among associated UEs, and, finally, the LBT

by the UE has to be succeeded in order for the scheduled UL transmission to be executed. Although the detailed operation will have to be specified, the autonomous UL access has several advantages over scheduled UL access. That is, any UE can perform LBT and can transmit on the UL whenever it has data to transmit, as the UL transmission is not limited by a UL grant. The autonomous UL access will also naturally well-coexist with Wi-Fi as the UE behavior, in terms of channel access, is not different from Wi-Fi stations.

References

- ETSI EN 301 893 (2017) Harmonized European Standard, Broadband Radio Access Networks (BRAN); 5 GHz high performance RLAN, V2.1.0
- IEEE Std 802.11TM-2012, Part 11: wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, 2012
- Jeon J, Niu H, Li Q, Papathanassiou A, Wu G (2015) LTE with listen-before-talk in unlicensed spectrum. In: IEEE international conference on communications, London
- Kwon H-J, Jeon J, Bhorkar A, Ye Q, Harada H, Jiang Y, Liu L, Nagata S, Ng BL, Novlan T, Oh J, Yi W (2017) Licensed-assisted access to unlicensed spectrum in LTE release 13. *IEEE Commun Mag* 55(2):201–207
- RP-151045 (2015) New work item on licensed-assisted access to unlicensed spectrum. Ericsson, Huawei, Qualcomm, Alcatel-Lucent, 3GPP TSG RAN Meeting #68
- RP-152272 (2015) New work item on enhanced LAA for LTE. Ericsson, Huawei, 3GPP TSG RAN Meeting #70
- RP-170828 (2017) New SID on NR-based access to unlicensed spectrum. Qualcomm
- RP-170848 (2017) New work item on enhancements to LTE operation in unlicensed spectrum. Nokia, Ericsson, Intel, Qualcomm, 3GPP TSG RAN Meeting #75
- Technical Report 36.889, Study on licensed-assisted access to unlicensed spectrum. 3GPP, May 2015

Certified Computation

- [Verifiable Cloud Computing](#)

Channel Access Methods

- [Multiple Access Techniques](#)

Channel Access Techniques

- [Multiple Access Techniques](#)

Channel Acquisition

- [Pilot Reuse for Massive MIMO](#)

Channel Delay Spread

- [Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System](#)

Channel Equalization for Block Transmission Schemes

- [Frequency-Domain Equalization in OFDM and Single-Carrier Modulation](#)

Channel Estimation

- [Parameter Estimation Exploiting WSNs with Arbitrary Topology Geometry](#)

Channel Impulse Response

- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

Channel Impulse Response (CIR)

- [Wireless Fading Channel Model for 5G and Future](#)

Channel Modeling of Microscale Terahertz Communication

Chong Han and Yi Chen
University of Michigan-Shanghai Jiao Tong
University Joint Institute (UM-SJTU JI),
Shanghai Jiao Tong University, Shanghai, China

Synonyms

[Terahertz channel modeling](#)

Definition

Channel modeling is the characterization of EM wave propagation in a channel, in the forms of channel impulse response and channel transfer function. The channel impulse response is the received signal that propagates through a wireless channel when the transmitted signal is an impulse, while the channel transfer function is the Fourier transformation of the channel impulse response. The channel model depends on distance, frequency, time, and propagation medium.

Historical Background

Terahertz band communication (0.1–10 THz) is envisioned as a key wireless technology to satisfy the growing demand for wireless Terabit-per-second (Tbps) links, by alleviating the spectrum scarcity. Channel modeling that characterizes the THz band is the fundamental of the designs for physical layer and data link layer in the THz communication systems. J.M. Jornet modeled the line-of-sight (LoS) THz channel in nanonetworks and detailedly analyzed the effect of molecular absorption (Jornet and Akyildiz 2010). S. Priebe introduced a stochastic spatiotemporal model has been for ultra-broadband THz indoor radio channels at 0.3 THz (Priebe and Kurner 2013). C. Han developed a ray-tracing based multipath channel for indoor scenario over 0.1–1 THz and a three-

dimensional end-to-end THz channel model (Han et al. 2015; Han and Akyildiz 2017). A hybrid channel modeling approach that combines ray tracing and statistical channel model is proposed for THz indoor channel modeling (Priebe et al. 2011).

Foundations

Terahertz band (0.1–10 THz) communication is envisioned as a key technology to support ultra-broadband wireless systems for beyond 5G. For realization of efficient wireless communication networks in the THz band, it is imperative to develop channel models which can accurately and efficiently characterize the THz spectrum peculiarities, which could be summarized as:

- Frequency-selective molecular absorption effect
- High free-space path loss
- High reflection and scattering loss
- Weak diffraction ability
- Extremely narrow beamwidth

In general, methods for channel modeling in wireless communications could be categorized as deterministic (Han et al. 2015), statistical (Priebe and Kurner 2013), and hybrid approaches (Thiel and Sarabandi 2008). The deterministic methods show high accuracy while suffering from high computational complexity. The computational complexity of statistical models is much lower than deterministic models. By the statistics of the channel properties, statistical models are adaptable to a range of propagation environment. However, the main drawback arises from the consistency with the real environment. To combine the benefits of these two methods, hybrid channel modeling balances the accuracy and low complexity (Han and Chen 2018).

Deterministic Channel Modeling

Deterministic channel models accurately model the wave propagation based on the electromagnetic wave theory. Deterministic approach is site-

specific, requiring detailed geometric information of propagation environment, properties of materials, and spatial positions of the transmitter and receiver. In general, this approach provides good consistency between the simulation results and the measurements.

Ray tracing (RT) has emerged as a popular technique for the analysis of site-specific scenarios, due to its ability to analyze very large structures with reasonable computational complexity (Han et al. 2015; Chen and Han 2018). As illustrated in Fig. 1, the radio signal transmitted from a fixed source encounters multiple objects in the environment, forming LoS, reflected, diffracted, or scattered arrival rays to the receiver. For each of these geometric paths, the RT techniques analyze the propagation of electromagnetic waves by representing the wave fronts as simple particles. In addition, the reflection, diffraction, and scattering effects are approximated by geometric optics, which becomes more accurate due to the stronger corpuscular property in the THz band. A complete and generic multipath channel model based on the RT method for the entire THz band has been established in Han et al. (2015). Furthermore, elevation plane is involved to develop a three-dimensional (3D) end-to-end channel model for the THz spectrum in Han and Akyildiz (2017)

Full-wave (FW) method, such as finite-difference time-domain (FDTD), method of

moments (MoM), and finite element method (FEM), is a numerical analysis technique that directly solves Maxwell's equations. As a main benefit, the full-wave method has the ability to resolve the small and complex scatterers and rough surface. However, this method suffers from very high computational consumption for high frequency with small wavelength and the object with complex geometry for the THz propagation.

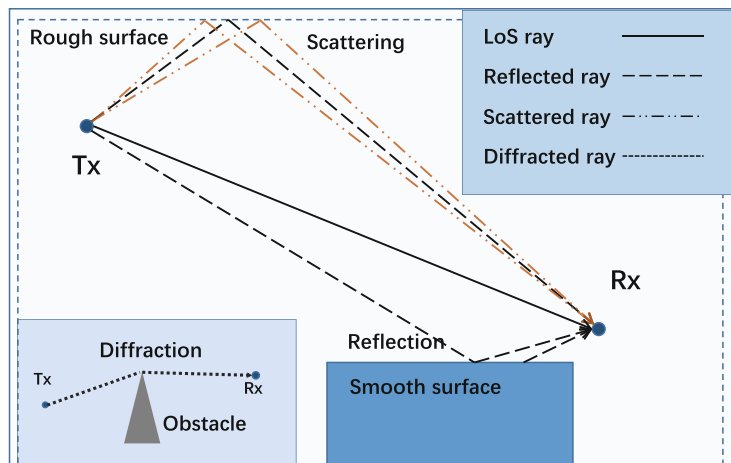
Statistical Channel Modeling

Although deterministic methods provide accurate channel modeling results, they require detailed geometric knowledge of the propagation environment and suffer from high computational complexity. By contrast, statistical methods could be invoked to model the channel based on empirical channel measurements.

For narrowband channels, the channel is characterized by large-scale fading model and small-scale fading model. The large-scale fading model characterizes path loss and shadowing. The small-scale fading model characterizes the effect of multipaths. In small-scale fading models, the received signal is a superposition of multipaths, and therefore its magnitude and phase can be regarded to follow a certain distribution. For example, Rayleigh fading model assumes that the magnitude of a signal follows a Rayleigh distribution, which is valid in heavily built-up urban environments range with no dominant line-of-sight path.

Channel Modeling of Microscale Terahertz Communication, Fig. 1

The ray-tracing method for deterministic channel modeling



For wideband THz channels, the multipaths are resolvable in the time domain. Therefore, a *statistical impulse response model* is developed where tap delay line formulas are established to characterize the wireless propagation. This approach specifies statistical distributions on the multipath parameters including direction-of-departure/direction-of-arrival (DoD/DoA), time-of-arrival (ToA), and complex amplitudes. Furthermore, based on the clustering observations of arrival rays, a cluster-based impulse response model considers the grouping effect of the multipath components, which have closely distributed ToA and DoA/DoD (Akdeniz et al. 2014). In particular, the well-known Saleh-Valenzuela (SV) model assumes ToA of clusters and intra-cluster rays follow Poisson processes, while the arrival time for each cluster is exponentially distributed.

Hybrid Channel Modeling

On one hand, deterministic channel modeling shows high accuracy with high time and resource consumption. On the other hand, statistical channel modeling has low computational complexity with lack of accuracy. Therefore, an interesting and promising direction is to develop *hybrid methods*, by strategically combining the benefits from two or more individual approaches. Currently, there are three kinds of hybrid approaches:

- **Stochastic Scatterer-placement and RT Hybrid Approach (SSRTH).** In this approach scatterers are stochastically placed, while the multipath propagation is traced and modeled based on RT techniques in a deterministic fashion. Based on the SSRTH approach, geometry-based stochastic channel models (GSCMs) are established (Molisch 2004; Kim and Zajić 2016). In the GSCMs, scatterers are generally assigned randomly on the basis of predetermined statistical distributions. As illustrated in Fig. 2, single-bounce and multi-bounce scattering are considered, and scatterers are grouped into

clusters. Next, simplified RT are then used to track and capture the propagation of the waves. The space-time parameters are then recorded for each path to construct the complete multipath channel impulse response.

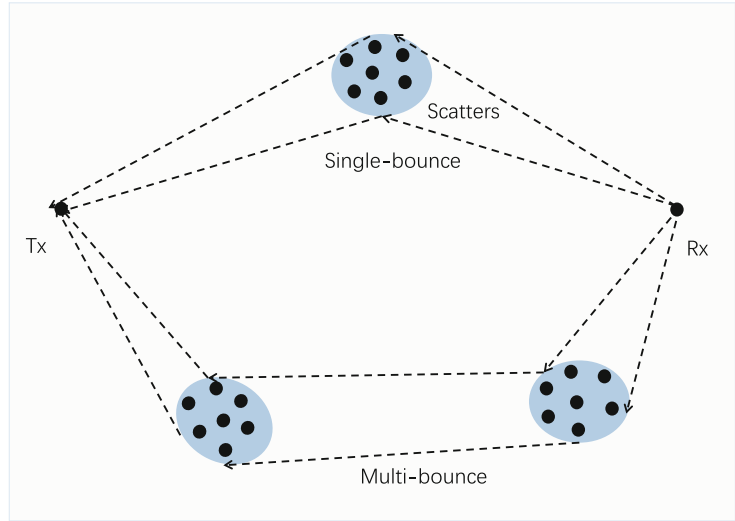
- **RT-FW Hybrid Modeling.** The principle of this method is to use FW to study regions close to complex discontinuities where ray-based solutions are not sufficiently accurate and to use RT to trace the rays out of the FW regions. This RT-FW hybrid technique is expected to provide very high accuracy in channel modeling. However, a critical challenge is to smoothly transit between the RT and full-wave methods and compute the boundary results.
- **Statistical and RT Hybrid (SRH) Approach.** Statistical impulse response channel modeling has high efficiency for characterizing the multipaths but lacks for accuracy. Therefore, a statistical and RT hybrid approach could be exploited. In particular, the dominant rays such as the LoS and the reflected are determined deterministically using RT techniques, while the other rays could be generated statistically. The main benefits include the high modeling accuracy, since the critical multipath components are computed deterministically, and low complexity since the very rich multipath effects are included using statistical modeling.

Ultra-massive MIMO Channel Modeling

Thanks to the very small wavelength at THz frequencies, hundreds and thousands of antennas can be packed in a small footprint. In the UM-MIMO channel with hundreds of antennas, the complexity increases exponentially if the deterministic channel modeling is applied to every individual sub-channel. The resulting computational and storage burden is very high (Ai et al. 2017). To address this problem, a *virtual point approximation approach* whose complexity is independent of the array size could be adopted to reduce the computing complexity. First, a point-to-point deterministic channel modeling is invoked to accurately capture the propagation between the virtual points at the transmitter and receiver as shown in Fig. 3.

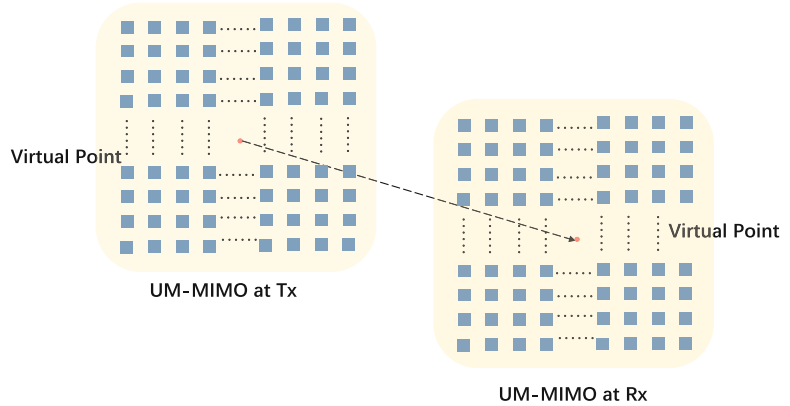
Channel Modeling of Microscale Terahertz Communication, Fig. 2

Geometry-based stochastic model with single-bounce and multi-bounce scattering



Channel Modeling of Microscale Terahertz Communication, Fig. 3

The virtual point approach for deterministic UM-MIMO channel modeling



Then, this propagation model is mapped to other antenna pairs, according to the relative spatial and angular separation in the antenna array.

Statistical MIMO models can be categorized into *matrix-based models* and *reference antenna-based models*. The matrix-based models focus on characterizing the properties of the complete channel transfer matrix. On the contrary, reference antenna-based models analyze the SISO propagation model between reference transmitting and receiving antennas first. Then, the complete channel transfer matrix is generated statistically, based on the developed reference single-antenna model.

The matrix-based models characterize the UM-MIMO channel matrix, in which each

entry of the matrix represents the multipath propagation of each sub-channel in the UM-MIMO communication system. The well-known matrix-based channel models are:

- **I.i.d. Rayleigh Model.** As a simplified consideration, the entries of the channel matrix are independent and identically distributed (i.i.d.) complex Gaussian random variables in the i.i.d. Rayleigh model.
- **Kronecker Model.** This model considers the correlation between sub-channels and assumes the correlation at transmitter and receiver are separable so that the correlation matrix is a Kronecker product of two correlation matrices at transmitter and receiver, respectively.

- **Virtual Channel Representation (VCR).** It gives an angular representation of the channel matrix by sampling the rays in a beam space; each entry of the channel matrix denotes the superposition of the rays in a spatial window (You et al. 2017).
- **Weichselberger Model.** As a combination of the Kronecker and VCR models, the Weichselberger model decorrelates the channel coefficients in the eigenspace rather than the beam space. Consequently, the Weichselberger model establishes a more generic framework of UM-MIMO channels without the restrictions of uniform linear array in VCR.

The reference antenna-based models synthesize the channel matrix by applying steering vectors to the SISO statistical model corresponding to two reference antennas at Tx and Rx, respectively (Wallace and Jensen 2002). The steering vector whose entry is phase shift relative to the reference antenna is based on the DoA/DoD and array configuration. The number of multipath components, path gains, ToAs, DoAs/DoDs, and the clustering properties are quantified parameters and randomly generated for the reference sub-channel. This kind of models is also called nongeometrical stochastic model since it doesn't relate to any geometry of the physical environment. However, since ToAs and DoAs/DoDs are considered unaltered over the array, the reference antenna-based models need to be carefully revised for wideband THz communications.

SSRTH in SISO hybrid channel modeling is also suggested to develop GSCMs to characterize the UM-MIMO propagation. This approach is naturally suitable for UM-MIMO channels because the spatial-temporal properties could be well captured in practice. Furthermore, the generated statistical map is independent of mobile station (MS) and shared among multiple links and time snapshots in the UM-MIMO propagation. This enables smooth time evolution and avoids channel discontinuity. The standard spatial channel models like 3GPP SCM, WINNER, and COST2100 channel models are the representative GSCMs at microwave frequencies. Besides, a

multifocal ellipse model has been established to capture the characteristics of UM-MIMO channel including spherical wave front effect and non-stationarity over the array (Wu et al. 2015).

Key Applications

The channel modeling helps evaluate the channel performance and the designs for physical layer and data link layer in the wireless communication systems. Deterministic channel modeling can replace channel sounding which is time-consuming in some cases. The statistical channel modeling is often used in the theoretical analysis, e.g., capacity, coverage, and bit error rate (BER).

Cross-References

- [Nanoscale Terahertz Communications](#)
- [Nanonetworks](#)

References

- Ai B, Guan K, He RS, Li JZ, Li GK, He D, Zhong ZD, Huq KMS (2017) On indoor millimeter wave massive MIMO channels: measurement and simulation. *IEEE J Sel Areas Commun* 35(7):1678–1690
- Akdeniz MR, Liu Y, Samimi MK, Sun S, Rangan S, Rappaport TS, Erkip E (2014) Millimeter wave channel modeling and cellular capacity evaluation. *IEEE J Sel Areas Commun* 32(6):1164–1179
- Chen Y, Han C (2018) Channel modeling and analysis for wireless networks-on-chip communications in the millimeter wave and terahertz bands. In: *IEEE INFOCOM 2018 Workshops*. IEEE, Honolulu
- Han C, Akyildiz IF (2017) Three-dimensional end-to-end modeling and analysis for graphene-enabled Terahertz band communications. *IEEE Trans Veh Technol* 66(7):5626–5634
- Han C, Chen Y (2018) Propagation modeling for wireless communications in the terahertz band. *IEEE Commun Mag* 56(6):96–101
- Han C, Bicen AO, Akyildiz IF (2015) Multi-ray channel modeling and wideband characterization for wireless communications in the Terahertz band. *IEEE Trans Wirel Commun* 14(5):2402–2412
- Jornet JM, Akyildiz IF (2010) Channel capacity of electromagnetic nanonetworks in the Terahertz band. In: *Proceedings of 2010 IEEE International Conference on Communications*, pp 1–6, Cape Town

- Kim S, Zajić A (2016) Statistical modeling and simulation of short-range device-to-device communication channels at sub-THz frequencies. *IEEE Trans Wirel Commun* 15(9):6423–6433
- Molisch AF (2004) A generic model for MIMO wireless propagation channels in macro- and microcells. *IEEE Signal Process Lett* 52(1):61–71
- Priebe S, Kurner T (2013) Stochastic modeling of THz indoor radio channels. *IEEE Trans Wirel Commun* 12(9):4445–4455
- Priebe S, Jacob M, Kuerner T (2011) AoA, AoD and ToA characteristics of scattered multipath clusters for THz indoor channel modeling. In: *Proceedings of 17th Europe an Wireless 2011 – Sustainable Wireless Technologies*, Vienna, pp 1–9
- Thiel M, Sarabandi K (2008) A hybrid method for indoor wave propagation modeling. *IEEE Trans Antennas Propag* 56(8):2703–2709
- Wallace JW, Jensen MA (2002) Modeling the indoor MIMO wireless channel. *IEEE Trans Antennas Propag* 50(5):591–599
- Wu S, Wang CX, Haas H, Aggoune EHM, Alwakeel MM, Ai B (2015) A non-stationary wideband channel model for massive MIMO communication systems. *IEEE Trans Wirel Commun* 14(3):1434–1446
- You L, Gao X, Li GY, Xia XG, Ma N (2017) BDMA for millimeter-Wave/Terahertz massive MIMO transmission with per-beam synchronization. *IEEE J Sel Areas Commun* 35(7):1550–1563

Channel Reciprocity

- [Wireless Key Establishment](#)

Charging Navigation

- [EV Charging Scheduling](#)

Chemical Reaction

- [Reaction-Diffusion Channels](#)

City Big Data

- [Urban Big Data](#)

Clock Synchronization of Nanomachines

- [Synchronization of Nanomachines](#)

Closed-Circuit Television (CCTV)

- [Interference Mitigation in Wireless Communications-Based Train Control \(CBTC\), Cognitive Control Approach](#)

Cloud Computing

- [Data Center Network in Cloud Computing](#)
- [Private Truth Discovery in Cloud-Empowered Crowdsensing Systems](#)

Cloud Storage Security

- [Distributed Storage Security](#)

Clustering

- [Application of Machine Learning in Wireless Sensor Network](#)

Cognitive Heterogeneous Networks

Lei Xu
Nanjing University of Science and Technology,
Nanjing, China

Synonyms

Heterogeneous wireless access medium with cognitive radio; Heterogeneous wireless networks based on cognitive radio

Definition

In cognitive heterogeneous wireless networks, the available spectrum resources for each network are dynamic and limited. It is difficult to provision quality of service (QoS) to secondary mobile terminals (MTs). Hence, integrating cognitive heterogeneous wireless networks can help to provide various classes of service to secondary MTs and to support seamless secondary MT roaming. In order to guarantee QoS for MTs, multi-homing technology can be applied, where the data stream from an MT is split into multiple sub-streams, and transmitted over multiple networks by different radio interfaces simultaneously (Xu et al. 2017a,b).

Historical Background

In the fifth generation (5G) mobile communication systems, heterogeneous wireless networks will use various wireless access technologies, e.g., macrocells providing low-to-medium rate services with a large coverage area, while microcells and Wi-Fi supporting high rate services in hotspots. In each wireless network, MTs do not always occupy all spectrum resources, and there exist vacant spectrum holes. Due to the spectrum resource scarcity, cognitive radio can be applied in each wireless network to utilize the vacant spectrum holes (Gunawardena and Zhuang 2011; Xu and Nallanathan 2016). This leads to cognitive heterogeneous wireless networks, where the coverage areas of multiple cognitive wireless networks overlap and secondary MTs can opportunistically utilize the temporary spectrum holes (Zhang et al. 2015). There are international standards, e.g., IEEE 802.11af, IEEE 802.19 TG 1, IEEE 802.22, and LTE-U, in place to promote the development of cognitive wireless networks (Xu et al. 2017c).

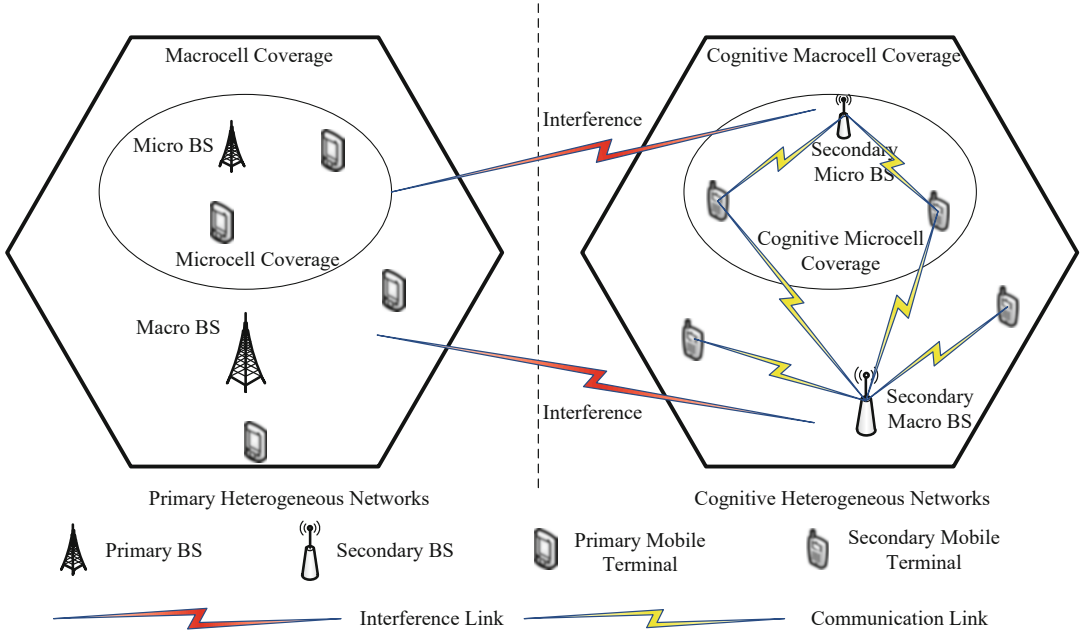
In cognitive heterogeneous networks, resource allocation is very important (Xu et al. 2017a,b). In Xu et al. (2017a), a call admission control algorithm based on inter-network cooperation is proposed via a Stackelberg game framework

for cognitive heterogeneous networks. The call admission control problem is subject to the variable bandwidth rate traffic, network service selection, feasible subchannel allocation, and call blocking probability. The call admission control algorithm is based on spectrum price at primary heterogeneous networks and subchannel allocation price and network selection at cognitive heterogeneous networks. In order to determine the call blocking probability, a probability upper bound of exceeding the maximum admission number for secondary MTs is analyzed based on M/M/ ∞ model. Then, the subchannel allocation price and network selection are designed via the dual decomposition method, and the vacant spectrum price is determined with Bertrand game theory. Finally, a call admission control algorithm is proposed.

In Xu et al. (2017b), a video packet scheduling framework with stochastic QoS is proposed for cognitive heterogeneous networks based on inter-network cooperation. The video packet scheduling is subject to constraints in the available energy at each call for secondary MT, the time-varying channel state information at different interfaces, the total interference power, the target call duration, and the video characteristics. The objective function maximizes the minimum lower bound of video quality. In order to solve the above video packet scheduling problem with stochastic QoS guarantee, a video packet scheduling scheme based on forward-auction theory is proposed. Then, the cumulative distribution function for video quality is analyzed. Finally, the power allocation scheme to maximize the minimization lower bound of video quality among different secondary MTs is presented.

Foundations

Consider a geographical region with $\mathcal{N} = \{1, 2, \dots, N\}$ primary wireless access networks, based on different wireless access technologies and operated by different service operators. In network n , there is a set, $\mathcal{S}_n = \{1, 2, \dots, S_n\}$, of base stations (BSs). Corresponding to network n BS s , there is a cognitive wireless



Cognitive Heterogeneous Networks, Fig. 1 Cognitive heterogeneous networks

network n BS s . The cognitive heterogeneous wireless networks are shown in Fig. 1. There is a set, $\mathcal{M} = \{1, 2, \dots, M\}$, of secondary MTs in the geographical region, and $\mathcal{M}_{ns} = \{1, 2, \dots, M_{ns}\} \in \mathcal{M}$ is a subset of the MTs, which reside in the coverage area of network n BS s . In primary network n BS s , the total spectrum is divided into K_{ns} subchannels, and each subchannel has the same bandwidth. Additionally, the transmission power at secondary BSs for downlink or at secondary MTs for uplink should be controlled to protect the transmission in primary networks according to the interference temperature model (Habachi et al. 2013). In the same cognitive network, interference mitigation is achieved by interference management schemes (Ahuja et al. 2015; Guler and Yener 2014). Using the multi-homing mechanism and multiple radio interfaces, each secondary MT can communicate with multiple secondary BSs simultaneously. In this work, we divide the spectrum into a number of subchannels. Primary networks use some subchannels, while cognitive networks utilize the vacant subchannels. The interference exists between primary networks and cognitive

networks due to the cross-channel interference, e.g., Zhang and Leung (2009) and Shaat and Bader (2012). Hence, the interference introduced by cognitive networks to primary networks should be controlled via the interference temperature model.

Consider K_{ns} time-varying subchannels licensed for primary network n BS s . The occupancy of each subchannel, $k \in \{1, 2, \dots, K_{ns}\}$, is modeled as a time-homogeneous discrete Markov process, denoted by s_{ns}^k , where $s_{ns}^k = 0$ means that the subchannel is idle and $s_{ns}^k = 1$ means that the subchannel is busy. Define α_{ns}^k (β_{ns}^k) as the probability that the licensed subchannel k at primary network n BS s changes from idle (busy) in the previous time slot to busy (idle) in the present slot. In cognitive heterogeneous networks, $1 - s_{ns}^k = 1$ indicates that the subchannel k is available for cognitive network n BS s ; otherwise $1 - s_{ns}^k = 0$ (Habachi et al. 2013). If $\rho_{nsm}^k = 1$, the subchannel k is allocated to cognitive network n BS s MT m . $\mathcal{K}_{ns}$ is the vacant subchannel set for cognitive network n BS s .

In a given service area, r , the subset of secondary MTs is $\mathcal{M}_r$. Besides its own home

network, each secondary MT can get service from other secondary BSs available at its location. Secondary MTs can be classified into two categories: network subscribers and network users. Each network subscriber uses its own home network, while each network user utilizes the other networks. Consequently, p_{nsm} is defined as a priority parameter, representing service priority for secondary network n in allocating its resources to secondary MT m via secondary BS s , where $p_{nsm} \in [0, 1)$ denotes for low-priority network users and $p_{nsm} = 1$ denotes for high-priority network subscribers. At each secondary MT, single-network service and multi-homing service are considered. The subset of secondary MTs using single-network service in the service area r is $\mathcal{M}_{1r}$, while the subset of secondary MTs using multi-homing service is $\mathcal{M}_{2r}$. $\mathcal{S}_r$ is the subset of all secondary BSs in the service area r . The set of secondary MTs in cognitive network n BS s , including MTs using either multi-homing service or single-network service, is denoted by $\mathcal{M}_{ns}$.

If secondary MT m uses single-network service in area r , the constraint is $\sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}_n} x_{nsm} = 1$, $m \in \mathcal{M}_{1r}$, $r \in \mathbf{R}$. x_{nsm} is a binary assignment variable, e.g., if secondary MT m communicates with cognitive network n BS s , $x_{nsm} = 1$; otherwise $x_{nsm} = 0$. $\mathbf{R}$ is the service area set. If secondary MT m uses multi-homing network service in area r and cognitive network n BS s serves in the service area r , the constraint is $x_{nsm} = 1$, $m \in \mathcal{M}_{2r}$, $r \in \mathbf{R}$.

Consider the video traffic at each secondary MT. In a geographic area, the video call arrivals are modeled as a Poisson process (Ghosh et al. 2011). Specially, in service area r , the arrival process for both new and handoff video calls is modeled by a Poisson process with arrival rate v_r . Additionally, the video call duration follows exponential distribution. The video call is variable bandwidth rate (VBR) traffic. For secondary MT m , a VBR call requires a bandwidth allocation within a minimum value, $B_m^{\min}$, and a maximum value, $B_m^{\max}$. In a given service area, if there are sufficient bandwidth resources, the VBR call will obtain the maximum required bandwidth, $B_m^{\max}$. When all secondary BSs reach their

capacity limitation, the bandwidth allocation for each VBR call is degraded toward the minimum required bandwidth, $B_m^{\min}$ (Xu et al. 2014).

Key Applications

Cognitive heterogeneous networks are fundamental network architecture in future wireless communication systems, e.g., 5G.

Cross-References

- 5G
- Cognitive Heterogeneous Networks
- Heterogeneous Wireless Access Medium with Cognitive Radio
- Heterogeneous Wireless Networks Based on Cognitive Radio

References

- Ahuja K, Xiao Y, van der Schaar M (2015) Efficient interference management policies for femto-cell networks. *IEEE Trans Wireless Commun* 14(9): 4879–4893
- Ghosh C, Roy S, Cavalcanti D (2011) Coexistence challenges for heterogeneous cognitive wireless networks in TV white spaces. *IEEE Wirel Commun* 18(4):22–31
- Guler B, Yener A (2014) Uplink interference management for coexisting MIMO femtocell and macrocell networks: an interference alignment approach. *IEEE Trans Wirel Commun* 13(4): 2246–2257
- Gunawardena S, Zhuang W (2011) Capacity analysis and call admission control in distributed cognitive radio networks. *IEEE Trans Wirel Commun* 10(9): 3110–3120
- Habachi O, El-azouzi R, Hayel Y (2013) A Stackelberg model for opportunistic sensing in cognitive radio networks. *IEEE Trans Wirel Commun* 12(5): 2148–2159
- Shaat M, Bader F (2012) Asymptotically optimal resource allocation in OFDM-based cognitive networks with multiple relays. *IEEE Trans Wirel Commun* 11(3):892–897
- Xu L, Nallanathan A (2016) Energy-efficient chance-constrained resource allocation for multicast cognitive OFDM network. *IEEE J Sel Areas Commun* 34(5):1298–1306

- Xu L., Li Y-P, Yang Y-W, Zhang X, Tang Z-M, Lan S-H (2014) Proportional fairness resource allocation scheme based on quantised feedback for multiuser orthogonal frequency division multiplexing system. *IET Commun* 8(16):2925–2932
- Xu L, Wang P, Li Q, Jiang Y (2017a) Call admission control with inter-network cooperation for cognitive heterogeneous networks. *IEEE Trans Wirel Commun* 16(3):1963–1973
- Xu L, Nallanathan A and Yang Y (2017) Video Packet Scheduling With Stochastic QoS for Cognitive Heterogeneous Networks. *IEEE Transactions on Vehicular Technology* 66(8):7518–7526
- Xu L, Nallanathan A, Song X (2017c) Joint video packet scheduling, subchannel assignment and power allocation for cognitive heterogeneous networks. *IEEE Trans Wirel Commun* 16(3):1703–1712
- Zhang Y, Leung C (2009) Resource allocation in an OFDM-based cognitive radio system. *IEEE Trans Commun* 57(7):1928–1931
- Zhang H, Chu X, Guo W, Wang S (2015) Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum. *IEEE Commun Mag* 53(3):158–164

Cognitive Radio

- [Cognitive Radio-Based Non-orthogonal Multiple Access](#)
- [Database-Based Spectrum Access](#)
- [Dynamic Spectrum Access Through Cognitive Radio Networking](#)
- [Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices](#)
- [Spectrum Resource Allocation in Cognitive Radio Networks](#)
- [Spectrum Sensing in Cognitive Radio Networks](#)

Cognitive Radio Networks

- [Cooperative Cognitive Radio Networks](#)

Cognitive Radio Sensor Networks

Ju Ren

School of Information Science and Engineering,
Central South University, Changsha, China

Synonyms

[Cognitive sensor networks](#)

Definitions

Cognitive radio sensor network (CRSN) is a kind of wireless sensor networks where sensor nodes and sink nodes are equipped with cognitive radio modules to opportunistically access spatiotemporally available channels to transmit sensed data (Akan et al. 2009). Different from traditional cognitive radio network, cognitive radio sensor network is more sensitive to the energy consumption in channel sensing and switching, which motivates the study of spectrum management focusing on improving the energy efficiency of power-limited sensor nodes.

Historical Background

Both of the concept proposals of wireless sensor network (WSN) and cognitive radio (CR) can be dated back to the 1990s or earlier (Mitola and Maguire 1999; Akyildiz et al. 2002). A WSN typically comprises a large number of battery-powered sensor nodes deployed in an interested area to perform environmental sensing and information converting and processing and then transmit their sensed data to sink nodes for further usage. With the low-cost and self-organized WSNs, humans are enabled to transfer various information data from the physical world to the cyber world in an

efficient manner and thereby promote the great fusion of cyber physical systems. Due to the power and cost limits of sensor nodes, short-range and low-power communication protocols operating on the unlicensed spectrum, such as Zigbee and Bluetooth, are widely adopted as the communication technologies for WSN to guarantee the energy efficiency of sensor nodes. However, with the explosion of mobile devices and services, the spectrum resources for wireless communications become increasingly scarce, making the data collection of WSNs face significant interference caused by nearby wireless applications. Thus, plenty of research efforts have been paid on addressing the energy-scarcity and spectrum-scarcity problems during recent years.

Cognitive radio has been proposed as a promising solution for alleviating spectrum-scarcity problem. It was first proposed by Joseph Mitola III in a seminar at KTH (the Royal Institute of Technology in Stockholm) in 1998 and published in 1999 (Mitola and Maguire 1999). The key idea of CR is to enable mobile devices to intelligently sense the ambient radio environment and dynamically configure the radio operation parameters to access the best vacant channel for opportunistic data transmission. After a few years' development, CR has experienced a big step forward from theory to practice. In Oct. 2004, the Institute of Electrical and Electronic Engineers (IEEE) initialized a special group, named IEEE 802.22 work group, to develop the international standards for CR. The standards were published in July 2011 and attracted a large number of CR test-beds and products in the past years.

The concept of CRSN was first proposed by Prof. Ozgur B. Akan in 2009 (Akan et al. 2009), with the aim of applying CR into WSN to address the spectrum-scarcity problem of WSN. As sensor nodes are sensitive to the energy consumption of spectrum sensing and channel switching, a key challenge of CRSN is to provide energy-efficient spectrum management strategies, including spectrum sensing, dynamic channel accessing, shar-

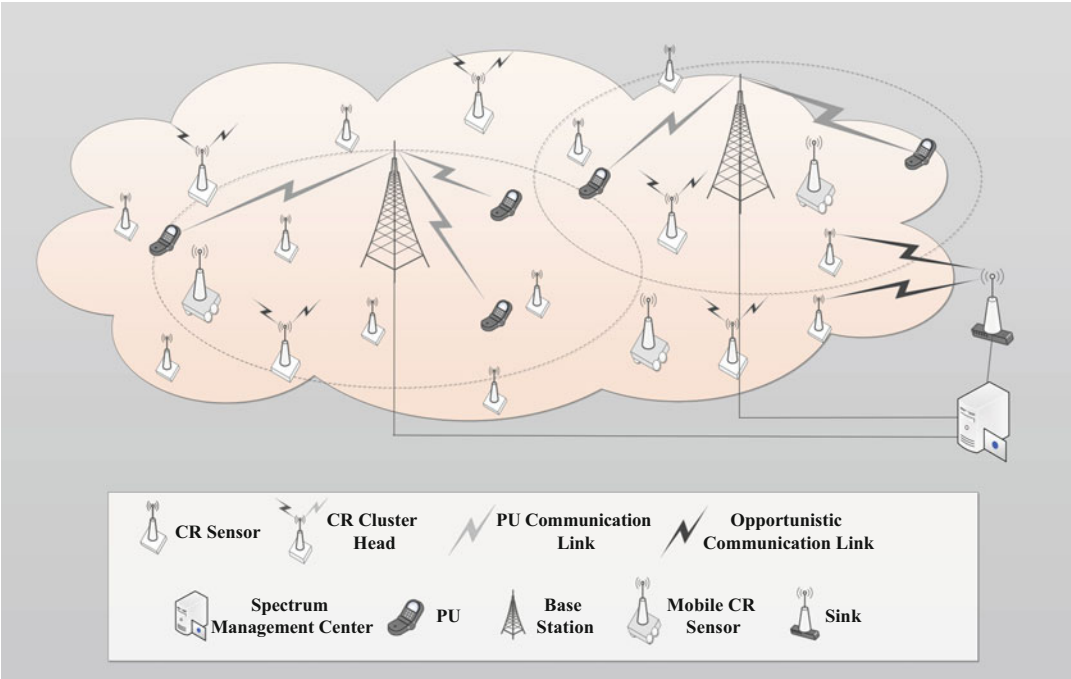
ing, and allocation, for guaranteeing the energy efficiency of the network.

Foundations

The architecture of CRSN is shown in Fig. 1, where a large number of distributed sensor nodes equipped with cognitive radio modules can opportunistically access vacant licensed channels to transmit their sensed data to the sink node via a multihop manner (Akan et al. 2009). The sink may also be equipped with CR capabilities to receive data via different channels. To efficiently manage the dynamic channel access in CRSN, there should be a default control channel for communicating the control information and coordinating the whole process.

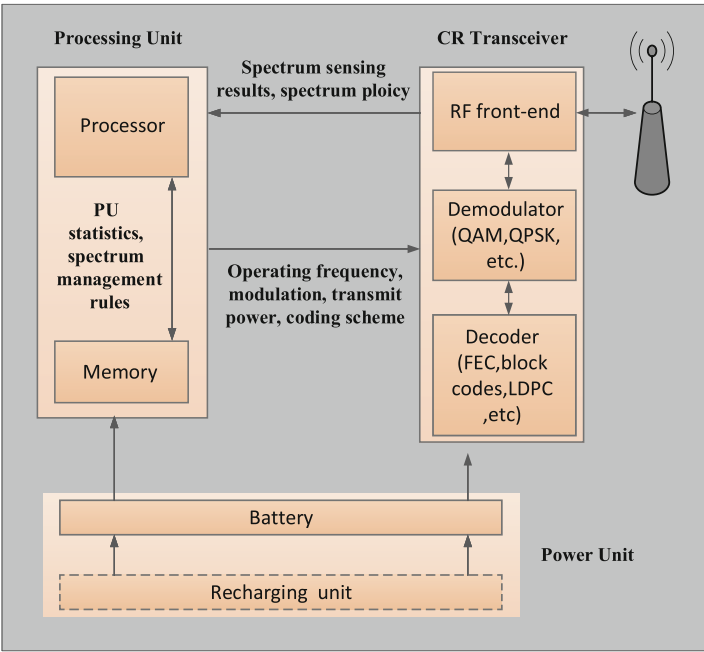
Figure 2 shows the hardware architecture of a cognitive radio sensor node (Akan et al. 2009). As shown in the figure, the main difference between the hardware structure of a classical sensor node and a CR sensor node is the cognitive radio transceiver. The cognitive radio unit enables the sensor nodes to dynamically adapt their communication parameters, such as carrier frequency, transmission power, and modulation. Cognitive radio sensor nodes also inherit the limitations of conventional sensor nodes in terms of power, communication, processing, and memory resources, which consequently restrict the features of cognitive radio.

Compared to traditional WSNs with fixed spectrum allocation, CRSN can benefit from many advantages brought by CR technology. Opportunistic channel access enables CRSN to achieve timely and bursty data transmission when the unlicensed spectrum is crowded. It also can improve the energy efficiency of data collection by avoiding data collision and retransmission over the highly interfered unlicensed spectrum. In addition, dynamic spectrum management significantly contributes to the efficient coexistence of spatially overlapping sensor networks or other



Cognitive Radio Sensor Networks, Fig. 1 The architecture of CRSN

Cognitive Radio Sensor Networks, Fig. 2 The hardware architecture of a cognitive radio sensor node



wireless applications in terms of communication performance and resource utilization.

Key Applications

CRSN can be widely applied in various scenarios where many overlapping wireless applications coexist making the unlicensed spectrum crowded. It can be well exploited to provide a spectrum- and energy-efficient data sensing and collection solution for delay-sensitive and bursty traffic applications. Some representative applications can be referred to E-healthcare systems, indoor multimedia sensing, and smart city applications.

Future Directions

As an emerging solution to address the spectrum scarcity problem in the IoT area, CRSN is still in its infancy and faces many obstacles hindering the flourish of its applications. Listed below are some potential future research directions:

- Distributed Spectrum Resource Management for Large-Scale CRSNs:* In large-scale CRSNs, some sensor nodes have to forward a significant amount of sensed data to the sink node. In order to improve the network lifetime and performance, multiple sink nodes or mobile sink nodes are usually deployed, accompanying with dynamic routing schemes, for balanced energy consumption. It consequently makes the spectrum resource management increasingly complicated and challenging. For large-scale CRSNs, centralized mechanisms/algorithms can no longer meet the QoS requirements in terms of processing delay and scalability. However, from the spectrum management perspective, centralized mechanisms/algorithms can provide more accurate control for spectrum sensing and accessing than distributed solutions (Ren et al. 2016a,b). Therefore, how to design efficient distributed spectrum
- resource management for addressing this dilemma in large-scale CRSNs becomes an challenging but meaningful research direction.
- Cross-layer Design for Opportunistic Spectrum Access:* Cross-layer design has been proved as an effective solution to improve the network performance during the past decade. The core idea is to maintain the functionalities associated with the original layers but to allow coordination, interaction, and joint optimization of protocols crossing different layers. There have been a large number of existing works focusing on cross-layer design for both wired and wireless networks. However, since most of them are proposed for the networks with specific communication channels, they are not suitable for CRSNs with opportunistic spectrum access. It motivates researchers to devote more attention on the cross-layer design for CRSNs to improve the overall network performance.
- RF Energy Harvesting and Transfer in CRSNs:* The flourish of WSN and CRSN applications is highly impeded by the limited energy supply of battery-powered sensor nodes. Thus, an emerging energy harvesting technology enables sensor nodes to absorb energy from ambient RF signals, which shows strong potentials to power future CRSNs (Ren et al. 2018). By RF energy harvesting, sensor nodes have no need to add additional energy harvesting devices, such as solar photovoltaics and wind turbines, to get recharged. Meanwhile, since CRSN has the capability of cognitively deciding to access different spectrum bands, it can sense a channel with the strongest RF signals as the energy harvesting channel and let the charging sensor nodes access to improve the charging efficiency (Zhang et al. 2017). Furthermore, with such a promising technology, sensor nodes can transfer energy among each other via RF signals to alleviate the unbalanced energy consumption in CRSNs. Therefore, CR and RF energy harvesting can be viewed as two complementary technologies and arise a number of interesting research topics in RF-powered CRSN applications.

Cross-References

- [Cognitive Radio Networks](#)
- [Internet of Things](#)
- [Wireless Sensor Networks](#)

References

- Akan OB, Karli O, Ergul O (2009) Cognitive radio sensor networks. *IEEE Netw* 23(4):34–40
- Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. *Comput Netw* 38(4):393–422
- Mitola J, Maguire GQ (1999) Cognitive radio: making software radios more personal. *IEEE Pers Commun* 6(4):13–18
- Ren J, Zhang Y, Ye Q, Yang K, Zhang K, Shen XS (2016a) Exploiting secure and energy-efficient collaborative spectrum sensing for cognitive radio sensor networks. *IEEE Trans Wirel Commun* 15(10):6813–6827
- Ren J, Zhang Y, Zhang N, Zhang D, Shen X (2016b) Dynamic channel access to improve energy efficiency in cognitive radio sensor networks. *IEEE Trans Wirel Commun* 15(5):3143–3156
- Ren J, Hu J, Zhang D, Guo H, Zhang Y, Shen X (2018) RF energy harvesting and transfer in cognitive radio sensor networks: opportunities and challenges. *IEEE Commun Mag* 56(1):104–110
- Zhang D, Chen Z, Ren J, Zhang N, Awad MK, Zhou H, Shen XS (2017) Energy-harvesting-aided spectrum sensing and data transmission in heterogeneous cognitive radio sensor network. *IEEE Trans Veh Technol* 66(1):831–843

Cognitive Radio-Based Non-orthogonal Multiple Access

Zhenyu Na, Mengshu Zhang, and Yuyao Wang
School of Information Science and Technology,
Dalian Maritime University, Dalian, Liaoning,
China

Synonyms

[Cognitive radio](#); [Non-orthogonal multiple access](#)

Definition

Cognitive Radio-based Non-orthogonal Multiple Access (CR-NOMA) is a derivation of NOMA. In CR-NOMA, a user with bad channel condition is viewed as the primary user (PU), while a user with good channel condition is regarded as the secondary user (SU), which is squeezed into the spectrum occupied by PU (Ding et al. 2016). CR-NOMA opportunistically serves one user on the condition that other users' Quality of Service (QoS) is guaranteed.

Historical Background

The advent of CR-NOMA is driven by the advantages of NOMA and CR. Recently, NOMA has attracted increasing attention because of its high spectral utilization. At the transmitter side, multiple users' signals can be transmitted simultaneously in time domain, frequency domain, or code domain by utilizing power domain multiplexing. At the receiver side, the multiplexed signals can be separated by using Successive Interference Cancellation (SIC). Thus, the resource block (time, frequency or code) is not solely occupied by one user but shared by multiple users. As an emerging technology to improve wireless spectrum utilization, CR has also attracted extensive attention because it allows SU opportunistically accessing to the spectrum initially licensed to PU (Lv et al. 2018). Intuitively, the combination of CR with NOMA can further improve radio resource utilization.

The radio resource management departments in different countries have something in common on wireless spectrum management. Specifically, the wireless spectrum allocation scheme in different countries often adopts the “fixed allocation” principle (Zhou et al. 2017), based on which the allowable spectrum is divided into several segments. Each segment is assigned to a user (or a group of users) and other users have no right to use it. The corresponding spectrum segment is called licensed band, while the user (or the

group of users) using the licensed band is called licensed user.

Since the popularization of smart terminals and the surge of mobile traffic stimulate the evolution of communications technology, as a kind of nonrenewable resources, the spectrum resource available is more scarce than ever before (Agiwal et al. 2016). A survey on spectrum usage from Federal Communications Commission (FCC) has pointed out that a large number of spectrum resources have different degrees of idleness in time and space dimensions even in some crowded bands (Kolodzy and Avoidance 2002). This is because only a small portion of spectrum is used frequently based on the “fixed allocation” principle (Zhou et al. 2017). In order to improve the utilization of the licensed spectrum as much as possible, various multiplexing techniques such as Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), and Multiple-Input Multiple-Output (MIMO) are put forward to support more users at the limited band (Wang and Liu 2011). However, they cannot fundamentally deal with the scarcity of spectrum resources caused by the “fixed allocation” principle.

In 1999, the concept of CR was first proposed by Dr. Mitola (Mitola and Maguire 1999). Then, Haykin defined CR from the communication point of view (Haykin 2005), pointing out that it is a form of intelligent wireless communication systems. He said that CR can automatically sense the usage of surrounding spectrum and use the idle spectrum without affecting the normal communications of licensed user. Generally speaking, the licensed user is referred to as PU, while the unlicensed user is known as SU or cognitive user. According to the definition of CR, SU should possess the abilities of spectrum sensing and reconfiguration simultaneously. On one hand, SU needs to detect whether there exists idle spectrum available. On the other hand, SU must hand over the spectrum immediately when PU needs to reuse it. It also needs to adjust the software/hardware parameters and working parameters according to the result of spectrum sensing.

So far, CR has been identified as a promising solution to spectrum scarcity and low spectrum utilization. Based on the key idea of Dynamic Spectrum Access (DSA), CR allows opportunistically utilizing idle spectrum without disturbing PU. So, multiple users or services can share the same spectrum segment, and then achieve the goal to avoid the high cost of spectrum resetting and improve spectrum utilization (Hu et al. 2018).

On account of the advantages of CR, it seems to be the perfect answer to address the challenges and requirements that the Fifth Generation (5G) of mobile communication systems poses (Tragos et al. 2013; Zhou et al. 2017).

In order to further ensure the sustainability of mobile communication services over the next decade, new technical solutions that can respond to future challenges and requirements need to be pondered and developed (Boccardi et al. 2014). For future Radio Access (RA) in the 2020 era, the mobile data traffic has been anticipated to undergo an exponential increase, e.g., at least 1000-fold in the 2020s compared with 2015. Therefore, the significant gains in the system capacity/efficiency and Quality of Experience (QoE) for users are required in the 2020 era. In cellular mobile communications, the design of RA technology is one important aspect to improve the system capacity in a cost-effective manner. In particular, the multiple access is a key element of RA technology. For example, in the Third Generation (3G) mobile communication systems, Wideband-CDMA (W-CDMA) and CDMA2000 and Direct Sequence-CDMA (DS-CDMA) are used whose receiver is based on simple single-user detection by using the Rake receiver. Orthogonal Multiple Access (OMA) based on Orthogonal Frequency Division Multiple Access (OFDMA) or Single Carrier-FDMA (SC-FDMA) is adopted in the 3.9G and 4G mobile communication systems (Razavi et al. 2017) such as Long Term Evolution (LTE) and LTE-Advanced (LTE-A) (Han et al. 2014; Higuchi and Kishiyama 2013). OMA is a reasonable and competent choice for achieving good system-level throughput performance in packet-domain services. However, the further enhancements in system efficiency and QoE,

especially at cell edge, are required in the 5G era. To accommodate such demands, NOMA has appeared as a promising candidate of multiple access scheme for the further enhancements of system performance. NOMA achieves user multiplexing typically in power domain that was not sufficiently utilized in previous generations of mobile communication systems. In NOMA, multiple users are multiplexed in power domain at the transmitter side and de-multiplexed at the receiver side by using SIC (Haykin 2005; Hu et al. 2018). From an information-theoretic perspective, NOMA with SIC is an optimal multiple access scheme according to the achievable multiuser capacity region (Tragos et al. 2013; Zhou et al. 2017; Rawat et al. 2016; Boccardi et al. 2014; Agiwal et al. 2016; Kolodzy and Avoidance 2002; Razavi et al. 2017; Higuchi and Kishiyama 2013). Compared with OMA, NOMA can contribute to the simultaneous enhancement of system efficiency and cell-edge user's experience. This is of particular importance to the future cellular systems where the channel conditions vary significantly among users due to the near-far effect.

Combining the advantages of CR and NOMA, CR-NOMA not only allows dynamically accessing to idle spectrum, but has the potential to accommodate more users (i.e., more SUs and/or more PUs) transmitting signals simultaneously according to NOMA principle, improving spectrum utilization, system throughput, and the fairness among users (Lv et al. 2018). These features are highly qualified for the applications oriented to 5G, such as Internet of Things (IoT) and low-power Machine-Type Communications (MTCs).

Foundations

CR is also called smart radio which has the abilities of intelligent learning and interaction with surrounding environment. Its core idea is to achieve DSA and spectrum sharing by spectrum sensing and learning ability. In CR, SU dynamically searches for "spectrum hole." When PU completes communications, SU should immediately detect the spectrum initially licensed by

PU, and then communicate on the idle spectrum. Whenever PU restarts communications, its signal should be detected by SU with an extremely high probability. Then, SU must release the spectrum or reduce transmit power to guarantee that there is no interference with PU. The reaction time of SU has strict regulations whose standard setting is one of the major works of CR.

In fact, spectrum utilization can be further improved because CR mainly focuses on the idle spectrum detection and the allocation to only one user (PU or SU), with less attention to the high-efficiency access scheme. For example, the traditional CR does not support the simultaneous communications between PU and SU on the same spectrum segment. Therefore, higher spectrum utilization is anticipated if the coexistence of PU and SU is allowed (Zheng et al. 2010). By introducing NOMA into CR, the opportunity of SU to access to the spectrum which is licensed to PU can be increased. Moreover, system throughput can be largely improved because the same spectrum is more efficiently utilized.

Though many NOMA schemes have been proposed for the future 5G RA techniques, such as Power Domain NOMA (PD-NOMA), Sparse Code Multiple Access (SCMA), Pattern Division Multiple Access (PDMA), Lattice Partition Multiple Access (LPMA), and Interleave Division Multiple Access (IDMA) (Ding et al. 2017), they are designed in the light of the similar concept that multiple users can be served in one orthogonal block (time slot, OFDMA subcarrier, or spreading code). As a typical case, PD-NOMA is elaborated.

In power domain, NOMA is realized by allocating different power levels to different users (Islam et al. 2016). Next, a simple example is introduced to elaborate the principle of NOMA. A Base Station (BS) serves two single-antenna users simultaneously on the same subcarrier. Here, we suppose that the weak user with bad channel condition (i.e., the user may be at cell edge) is PU and the strong user with good channel condition is SU (i.e., the user may be close to BS). We hope the QoE of PU can be satisfied. The channels of PU and SU are denoted by h_1 and h_2 . BS superimposes their signals by allocating

different power coefficients which are denoted by α_1 and α_2 . The key idea of NOMA is to allocate more power to PU (Zabetian et al. 2017). That means, if $|h_1| < |h_2|$, then $\alpha_1 > \alpha_2$ with $\alpha_1^2 + \alpha_2^2 = 1$. Thus, the signal of PU is strong at SU. In other words, SU receives a strong interference from PU because of more transmit power allocated to PU. In contrast, PU is quite less influenced by the inference from SU because of less transmit power allocated to SU. Therefore, PU decodes its own signal directly by treating the signal from SU as noise, and the achievable rate of PU is expressed as

$$R_1 = \log_2 \left(1 + \frac{|h_1|^2 \alpha_1^2}{|h_1|^2 \alpha_2^2 + \sigma^2} \right),$$

where σ^2 is the normalized noise variance. Then, SU performs SIC. Specifically, it decodes the signal of PU, and then subtracts this signal from its observation. Finally, it decodes its own signal. The achievable rate of SU is formulated as

$$R_2 = \log_2 \left(1 + \frac{|h_2|^2 \alpha_2^2}{\sigma^2} \right).$$

In contrast, for the case of two users in OMA, we suppose the allowable spectrum is divided into two parts denoted by β_1 and β_2 , while $\beta_1 + \beta_2 = 1$. For simplicity, we only consider the unit spectrum (1 Hz). Thus, the achievable rate of PU and SU are expressed as

$$\begin{aligned} R_1 &= \beta_1 \log_2 \left(1 + \frac{\alpha_1^2 |h_1|^2}{\sigma^2} \right) \\ R_2 &= \beta_2 \log_2 \left(1 + \frac{\alpha_2^2 |h_2|^2}{\sigma^2} \right) \end{aligned}$$

Since NOMA adjusts the users' throughputs by changing their power allocation ratio, the throughput and fairness between these two users can be improved and optimized. On the condition that the two users have different Signal-to-Noise Ratios (SNRs), it can be easily proved that the throughput based on NOMA is higher than that based on OMA. In other words, NOMA is more effective in dealing with different SNRs.

NOMA allocates more power to PU with bad channel condition and less power to SU with good channel condition to guarantee the fairness between PU and SU. However, it cannot strictly provide the targeted QoS to PU. Fortunately, CR-NOMA is able to deal with this situation.

CR-NOMA aims at treating NOMA as a special case of CR, where the power allocations to PU and SU conditioned on their predefined QoS requirements (Ding et al. 2016). For example, in the downlink transmission with NOMA scheme, enough power should be allocated to PU with bad channel condition to strictly satisfy a data rate threshold as follows

$$\log_2 \left(1 + \frac{|h_1|^2 \alpha_1^2}{|h_1|^2 \alpha_2^2 + \frac{1}{\text{SNR}}} \right) \geq R_T$$

where R_T denotes the target data rate. If there is any power left afterwards, SU can be served by the remaining power. CR-NOMA guarantees that PU will always be served by sufficient power to satisfy its QoS requirements. By using CR-NOMA, not only PU can be served with its targeted QoS, but also SU can be allowed to use this subcarrier. Therefore, the achievable system throughput can be increased on the premise of fairness between PU and SU.

Key Applications

So far, CR has been applied to many fields. In the civil field, CR is adopted in Wireless Regional Area Network (WRAN), Adaptive Heuristic for Opponent Classification (AdHoc) network, Ultra-Wideband (UWB) system, Wireless Local Area Network (WLAN), Mesh network, and MIMO system. In the military field, CR can also be adopted since it can improve system throughput, spectrum management efficiency, electronic countermeasures, and system interconnection and interoperability capability.

As a multiple access technology in power domain, NOMA can greatly improve spectrum utilization and device access capability. It can obtain robust system gain without relying

on the Channel Quality Indicator (CQI) or Channel State Information (CSI). Furthermore, it can reduce the network transmission delay and terminal consumption by simplifying transmission scheduling. The combination of NOMA and other communications technologies can further improve the system performance.

In 5G networks, there are many typical performance requirements, such as high system throughput, low latency, large-scale connections, high spectrum management efficiency, and so on. As the combinational technology of NOMA and CR, CR-NOMA can evidently improve a certain metric on the premise of guaranteeing other metrics. Because CR-NOMA integrates the advantages of CR and NOMA, it can be applied to most of application fields of NOMA and CR, especially in the 5G networks. CR-NOMA can provide services for multiple users at the same time and provide efficient spectrum utilization. Recently, the industry is occupied in the effort to incorporate CR-NOMA into 5G network. The advances of LTE-A and digital television standards also show that CR-NOMA will become a part of the future generation of wireless networks.

Future Directions

As a new form of non-orthogonal access technology, CR-NOMA is receiving widespread attention. A lot of research on CR-NOMA and its derivative techniques oriented to 5G system is in progress. We only list some possible directions and challenges face by CR-NOMA in the future, including the following:

- CR-NOMA needs to further split the difference between the system throughput and the cell-edge user throughput. To deal with it, the more effective power allocation and user matching algorithms are necessary.
- The combination of CR-NOMA and MIMO technology can support more users at the same time, and further improve the system throughput. Therefore, the joint design of precoding

and power allocation should be further studied.

- Since cooperative relaying technique can improve the signal reception, it can be applied to CR-NOMA to further improve the communication performance of PU. However, this work is still in its infancy and needs further investigation.

Cross-References

- [Database-Based Spectrum Access](#)
- [Game Theory-Assisted Cooperative Spectrum Access](#)
- [MAC in Cognitive Radio Networks](#)
- [Spectrum Resource Allocation in Cognitive Radio Networks](#)

References

- Agiwal M, Roy A, Saxena N (2016) Next generation 5G wireless networks: a comprehensive survey. *IEEE Commun Surv Tutor* 18(3):1617–1655
- Boccardi F, Heath RW, Lozano A et al (2014) Five disruptive technology directions for 5G. *IEEE Commun Mag* 52(2):74–80
- Ding Z, Fan P, Poor HV (2016) Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. *IEEE Trans Veh Technol* 65(8):6010–6023
- Ding Z, Lei X, Karagiannis GK, Schober R, Yuan J and Bhargava VK (2017) A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. *IEEE J Sel Areas Commun* 35(10):2181–2195
- Han S, Cih-Lin I, Xu Z et al. (2014) Energy efficiency and spectrum efficiency co-design: from NOMA to network NOMA. *IEEE COMSOC MMTC E-Letter* 9(5):21–24
- Haykin S (2005) Cognitive radio: brain-empowered wireless communications. *IEEE J Sel Areas Commun* 23(2):201–220
- Higuchi K, Kishiyama Y (2013) Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink. In: *Vehicular technology conference (VTC Fall)*, 2013 IEEE 78th. IEEE, pp 1–5
- Hu F, Chen B, Zhu K (2018) Full spectrum sharing in cognitive radio networks toward 5G: a survey. *IEEE Access* 6:15754–15776
- Islam SMR, Avazov N, Dobre OA, Kwak K et al. (2016) Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. *IEEE Commun Surv & Tutor* 19(2):721–742

- Kolodzy P, Avoidance I (2002) Spectrum policy task force. Federal Commun Comm, Washington, DC, Rep. ET Docket, 40(4): 147–158
- Lv L, Yang L, Jiang H, Luan T.H, Chen J et al. (2018) When NOMA meets multiuser cognitive radio: opportunistic cooperation and user scheduling. *IEEE Trans Veh Technol* 67(7):6679–6684
- Mitola J, Maguire GQ (1999) Cognitive radio: making software radios more personal. *IEEE Pers Commun* 6(4):13–18
- Rawat P, Singh KD, Bonnin JM (2016) Cognitive radio for M2M and internet of things: a survey. *Comput Commun* 94:1–29
- Razavi R, Dianati M, Imran MA (2017) Non-orthogonal multiple access (NOMA) for future radio access. In: *5G mobile communications*. Springer, Cham, pp 135–163
- Tragos EZ, Zeadally S, Fragkiadakis AG et al (2013) Spectrum assignment in cognitive radio networks: a comprehensive survey. *IEEE Commun Surv Tutor* 15(3):1108–1135
- Wang B, Liu KJR (2011) Advances in cognitive radio networks: a survey. *IEEE J Sel Top Sign Proces* 5(1): 5–23
- Zabetian N, Baghani M, Mohammadi A (2017) Rate optimization in NOMA cognitive radio networks. In: *International symposium on telecommunications*. IEEE, pp 62–65
- Zheng G, Ma S, Wong KK et al (2010) Robust beamforming in cognitive radio. *IEEE Trans Wirel Commun* 9(2):570–576
- Zhou Z, Guo D, Honig ML (2017) Licensed and unlicensed spectrum allocation in heterogeneous networks. *IEEE Trans Commun* 65(4):1815–1827

Cognitive Sensor Networks

► Cognitive Radio Sensor Networks

Collaborative Beamforming in Wireless Sensor Networks

Xuecai Bao
Nanchang Institute of Technology, Nanchang,
China

Synonyms

[Distributed beamforming in wireless sensor networks](#)

Definitions

Collaborative Beamforming (CB) in Wireless Sensor Networks (WSNs) is a transmission technique of improving the energy efficiency and signal gain between collaborating nodes and the base station (BS) in one-hop transmission. The collaborative nodes in CB use the way of cooperative communication to form the high gain and directivity in the direction of the intended BS or sink node.

Historical Background

WSNs have been playing an increasing role in many application areas, such as agriculture, industry, environmental monitoring, and so on. From another point of view, due to the limited energy and transmission distance for the sensor node in WSNs, many research studies focus on the design of method or scheme for improving the energy efficiency and the network performance. The purpose is to reduce the energy consumption and prolong the network lifetime. In recent years, the CB provides a promising technique in increasing the transmission distance and improving the energy efficiency for WSNs. Ochiai et al. (2004) are the first present the concept of CB. The authors mainly focus on the analysis of the beampattern performance for the collaborative nodes randomly located at the given areas and the feasibility of CB communication in WSNs. In the same year, Barriac et al. (2004) proposed a similar concept called distributed beamforming in WSNs. This literature mainly discussed the realizability of the phase and frequency synchronization. Based on the feasibility of achieving phase synchronization, many works on CB in WSN have been proposed. These research studies can be divided into two classes, i.e., the statistical analysis of beampattern and the performance optimization of beampattern, respectively.

In terms of the statistical analysis of beampattern, these works focus on the analysis of beampattern performance for different random

distributions of sensor nodes. Ochiai et al. (2005) are the first to present the statistical performance of the CB beampattern for the uniformly distributed sensor nodes over a disk. Some conclusions can be drawn from the analyzed results, e.g., the narrow mainlobe can be obtained as the number of sensor nodes in the large disk increases. Also, when the deployment of sensor nodes is sparse enough, the directivity can approach N . Furthermore, Ahmed and Vorobyov (2009) analyze the performance of the Gaussian distribution of sensor nodes. The results show that the Gaussian distributed nodes have a lower directivity.

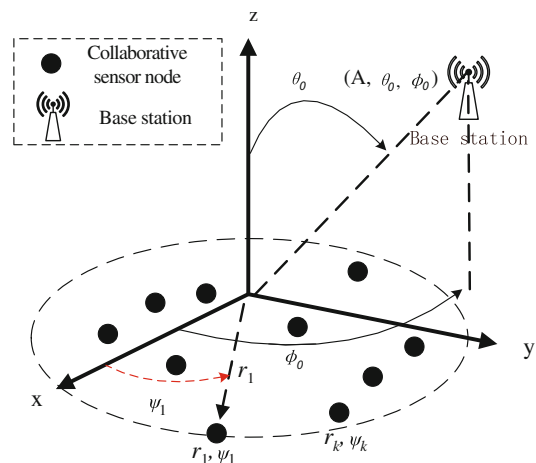
However, achieving the performance requires a large number of sensor nodes deploying at the disk areas, which is unrealistic for the practical WSN applications. Therefore, in order to improve the beampattern performance, the beampattern optimization of CB in WSN has attracted increasing attentions in recent years. The objectives of beampattern optimization mainly include minimizing the sidelobe level (Malik et al. 2009), maximizing the directivity (Jayaprakasam et al. 2014), minimizing the total power (Schedler and Kuehn 2014), and maximizing the network lifetime (Ahmed and Vorobyov 2011). Most of literatures focus on the minimizing sidelobes. The methods of minimizing sidelobes include the minimization of peak sidelobe (PSL) and sidelobe control in the direction of several unintended BS. Generally, the above optimization objectives are achieved via the coefficient perturbation and node selection. The coefficient refers to the amplitude of signal or the power transmission signal. The node selection needs to use the optimization method to select a subset of nodes from N collaborative nodes. Since the optimization problem is nonconvex, the heuristic algorithms are used for optimizing the beampattern of CB in WSN, such as particle swarm algorithm (PSO) (Malik et al. 2009), genetic algorithm (GA) (Wong et al. 2012), and so on. Moreover, due to high computational complexity of heuristic algorithms, Chen et al. (2016) proposed a node selection algorithm with a lower complexity, which is achieved based on cross entropy optimization (CEO).

The corresponding decentralized implementation of CEO is also presented. As the mentioned in Jayaprakasam et al. (2017), the future works need to consider the more practical characteristics, such as the nonideal antenna and channel, nonperfect synchronization, etc.

Foundations

The beampattern analysis of CB in WSNs needs the foundation of the knowledge of random antenna array. The more detailed description is given in Lo (1964) and Ochiai et al. (2005). To facilitate the understanding of CB in WSNs, some corresponding definitions and notions in CB need to be explained. The specific description is presented in (Fig. 1).

Array Factor Model: The array factor model is critical for the analysis of beampattern for CB in WSNs. Before presenting the array factor model, some assumptions need to be given. In WSNs, the N collaborating nodes are randomly deployed at circle areas and the polar coordinates are used to denote the location of each node. The corresponding coordinate of the k_{th} node is denoted by (r_k, ψ_k) . Similarly, the coordinates of the received BS is denoted by (A, ϕ_0, θ_0) . Furthermore, each node has location information of all nodes and



Collaborative Beamforming in Wireless Sensor Networks, Fig. 1 Collaborative beamforming scenario in WSNs

the perfect synchronization. The two dimensional model is considered, i.e., the sensor nodes and BS are on the same plane. Therefore, $\theta = \theta_0 = \frac{\pi}{2}$. Based on these assumptions, the array factor for the N nodes in WSNs can be written as

$$AF(\phi, \theta | \mathbf{r}) = \frac{1}{N} \sum_{k=1}^N e^{j\psi_k} e^{j\frac{2\pi}{\lambda} d_k(\phi, \theta)} \quad (1)$$

where $d_k(\phi, \theta) = \sqrt{(A^2 + r_k^2 - 2r_k A \sin\theta \sin(\phi - \psi_k))}$ is the distance between the k_{th} collaborating node and the received BS, and ψ_k is the initial phase of collaborating node $k \in 1, 2, \dots, N$. The initial phase can be obtained by the closed-loop method and the open-loop method (Ochiai et al. 2005). When the initial phase ψ_k is equal to $-\frac{2\pi}{\lambda} d_k(\phi, \theta)$, the maximum amplitude in direction of the intended BS (A, ϕ_0, θ_0) is obtained (As shown in Fig. 2).

Mainlobe and Sidelobes: The lobe including the maximum radiated power in beam pattern of CB is referred to the mainlobe. Moreover, the lobes that remove the main lobe in beam pattern of CB are sidelobes. In Fig. 2, the regions of the mainlobe and sidelobes are given.

Far-field Beam pattern: When the condition of far-field $A \gg r_k$ holds, i.e., the distance between the k_{th} node and BS is much larger than the r_k , the array factor is approximated by

$$AF(\phi, \theta | \mathbf{r}) = \frac{1}{N} \sum_{k=1}^N e^{j\frac{4\pi}{\lambda} r_k \left[\sin\left(\psi_k - \frac{\phi}{2}\right) \sin\frac{\phi}{2} \right]} \quad (2)$$

If $z_k = r_k \sin\left(\psi_k - \frac{\phi}{2}\right)$, the $AF(\phi | \mathbf{r}, \theta) = AF(\theta | \mathbf{z})$. The far-field beam pattern can be defined as $P(\theta | \mathbf{z}) = |AF(\theta | \mathbf{z})|^2$.

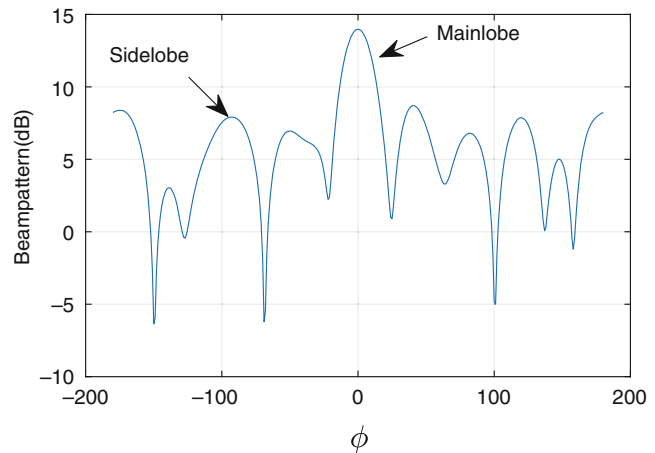
Average Beam pattern: The average beam pattern is defined by the statistical average of the far-field beam pattern, i.e., $P_{av}(\theta) = E_z\{P(\theta | \mathbf{z})\}$, where $\mathbf{z} = [z_1, z_2, \dots, z_N] \in (-\infty, +\infty)^N$.

Directivity: The directivity denotes the concentration level of radiated energy in the intended direction.

Synchronization: The perfect synchronization is important for CB communication in WSNs. The collaborative nodes compensate the phase offset such that the intended BS of receiving the signal from collaborative nodes can achieve the high gain. From Eq. (1), the initial phase provides the phase offset to implement CB. The methods of achieving synchronization are performed by the bit feedback from the intended BS. The process of bit feedback generally includes the following steps. First, each collaborating sensor node sends a test signal to the BS. When the BS receives the signal, the BS will broadcast the feedback information, such as channel state and signal amplitude, to all collaborative nodes. Finally, each collaborative node updates the coef-

Collaborative Beamforming in Wireless Sensor Networks, Fig. 2

Mainlobe and sidelobes in Beam pattern



ficient perturbation to achieve the synchronization. The more detailed studies of synchronization are given in Jayaprakasam et al. (2017).

Key Application

The CB technique in WSNs provides a lower energy consumption and longer transmission range, which has been widely used in various monitoring applications, such as agriculture, environment, volcano hazard monitoring. Specially, when the distance between sensor nodes and the BS is longer than 5 km Felici-Castell et al. (2017), CB in WSNs can achieve the long distance communication and power gain. In the future, as the improvement of both theoretical methods and practical implementation technologies, CB in WSNs will provide more the effective solution for more monitoring applications which require higher energy efficiency and longer distance communication.

Cross-References

- [Data Gathering in Wireless Sensor Networks](#)
- [Lens Antenna Array](#)

References

- Ahmed MF, Vorobyov SA (2009) Collaborative beamforming for wireless sensor networks with gaussian distributed sensor nodes. *IEEE Trans Wirel Commun* 8(2):638–643
- Ahmed MF, Vorobyov SA (2011) Power control for collaborative beamforming in wireless sensor networks. In: 2011 conference record of the forty fifth Asilomar conference on signals, systems and computers (ASILOMAR), IEEE, Pacific Grove, CA, pp 99–103
- Barriac G, Mudumbai R, Madhow U (2004) Distributed beamforming for information transfer in sensor networks. In: Proceedings of the 3rd international symposium on Information processing in sensor networks, ACM, Berkeley, CA, pp 81–88
- Chen JC, Wen CK, Wong KK (2016) An efficient sensor-node selection algorithm for sidelobe control in collaborative beamforming. *IEEE Trans Veh Technol* 65(8):5984–5994
- Felici-Castell S, Navarro EA, Pérez-Solano JJ, Segura-García J, García-Pineda M (2017) Practical considerations in the implementation of collaborative beamforming on wireless sensor networks. *Sensors* 17(2):237
- Jayaprakasam S, Rahim SKA, Leow CY, Yusof MFM (2014) Beamforming optimization in distributed beamforming using multiobjective and metaheuristic method. In: 2014 IEEE Symposium on Wireless Technology and Applications (ISWTA), Kota Kinabalu, IEEE, pp 86–91
- Jayaprakasam S, Rahim SKA, Leow CY (2017) Distributed and collaborative beamforming in wireless sensor networks: classifications, trends, and research directions. *IEEE Commun Surv Tutor* 19(4):2092–2116
- Lo Y (1964) A mathematical theory of antenna arrays with randomly spaced elements. *IEEE Trans Antennas Propag* 12(3):257–268
- Malik NNNA, Esa M, Yusof SKS (2009) Optimization of adaptive linear sensor node array in wireless sensor network. In: Microwave conference, 2009. APMC 2009. Asia Pacific, Singapore, IEEE, pp 2336–2339
- Ochiai H, Mitran P, Poor HV, Tarokh V (2004) Collaborative beamforming in ad hoc networks. In: Information theory workshop, 2004. IEEE, pp 396–401
- Ochiai H, Mitran P, Poor HV, Tarokh V (2005) Collaborative beamforming for distributed wireless ad hoc sensor networks. *IEEE Trans Signal Process* 53(11):4110–4124
- Schedler S, Kuehn V (2014) Resource allocation for distributed beamforming with multiple relays and individual power constraints. In: 2014 11th International Symposium on Wireless Communications Systems (ISWCS), Barcelona, IEEE, pp 1–5
- Wong CH, Siew ZW, Tan MK, Chin RKY, Teo KTK (2012) Optimization of distributed and collaborative beamforming in wireless sensor networks. In: 2012 fourth international conference on Computational Intelligence, Communication Systems and Networks (CICSyN), Mathura, IEEE, pp 84–89

Collaborative Cheating

- [Collusion](#)

Collaborative Consumption

- [Sharing Economics](#)

Collaborative Multipath Transmission

Yuanlong Cao, Li Zeng, Qinghua Liu, Gang Lei, Hao Wang, and Minghe Huang
Jiangxi Normal University, Nan Chang, China

Synonyms

Multiple network paths; Multipath transport protocol

Definitions

Multiple network paths can be established by the protocols on different network stack layers and can be used to send traffic data, enhance the connection reliability, achieve load balancing, etc.

Multipath transmission technology allows a network device equipped with multiple network interfaces (a.k.a a “multi-homed” device) to use multiple network paths for data transmission.

Collaborative multipath transmission is a promising technology toward collaboration multipathing by involving the full cooperation of all network elements (e.g., network nodes and protocol stacks) in multipath transmission.

Historical Background

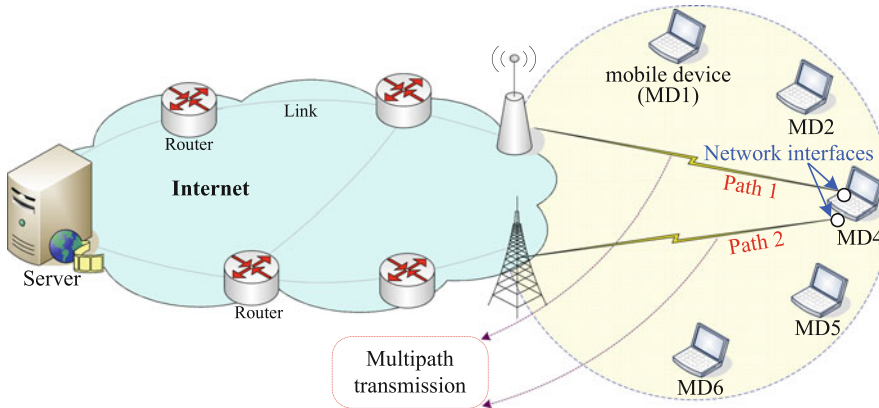
In the past several years, wireless networking technologies have undergone a rapid development. Significant advances in these wireless technologies have enabled today’s network devices to be equipped with multiple network interfaces (Sinky et al. 2019; Elgabli et al. 2019). Such multi-homed devices can exploit the connection diversity of multiple paths to possibly guarantee data transmission performance and protect against the single-path failures, supported by the emerging multipath transmission protocols (Li et al. 2016). Figure 1 shows a typical multipath transmission scenario which involves

two multi-homed devices (server and mobile device 4 (MD4)) that connect to each other through two paths (1 and 2). Such multipath transmission way can enable high-level quality of service provisioning for data delivery.

Current multipath transmission deployments exploit multiple paths to provide high-level transmission performance in two ways: (i) *Primary-alternative single-path transfer*. In this situation, the sender selects one path as the primary path, which is used for traffic data transmission. If the primary path becomes unavailable, the sender selects one of the alternate paths to perform continuous transmission of that data. (ii) *Concurrent multipath transfer*. In this situation the sender makes use of multiple network interfaces to perform concurrent transfer of data over all the available paths in the transmission session, achieving load balancing and increasing the throughput (goodput) performance.

Due to its attractive multipathing feature, multipath transmission technologies attract considerable attention and have resulted in many relevant Internet drafts, research publications, and industry applications. On the aspect of standardization, several IEEE and IETF working groups have concentrated their efforts on the standardization of multipath on different TCP/IP protocol stack layers (Lei et al. 2018; Devetak and Kapoor 2017; Stewart 2007; Ford et al. 2013; 802.1AX 2008; IEEE 802.1aq 2012; Eastlake et al. 2008, 2014). For example, the IETF has released two promising transport layer multipath transmission protocols, Stream Control Transmission Protocol (SCTP) (Stewart 2007) and Multipath Transmission Control Protocol (MPTCP) (Ford et al. 2013), in order to enable the concurrent use of multiple IP addresses/interfaces toward transport layer multipathing.

On the aspect of industry applications, several big high-tech companies are investing considerable amounts of research resources in the multipath technologies. For example, Samsung is dedicated to developing and deploying the Download Booster (Samsung 2018), which allows a multi-homed device to speed up the download of large files by using both the Wi-Fi and cellular connections simultaneously. Apple



Collaborative Multipath Transmission, Fig. 1 A typical multipath communication scenario between two multi-homed devices

Corporation applies the MPTCP technology to the iOS-based products (Apple 2018) in order to improve devices' network reliability and performance. Tessares company (Detal et al. 2017) tried to combine xDSL with LTE to create hybrid access networks, by using the promising MPTCP technology.

On the aspect of academic research, thousands of peer-reviewed publications revolving around multipath transmission technologies have been released in the past few years, covering the application, transport, network, data link, and physical layers of the TCP/IP protocol stack and involving a wide range of aspects such as packet scheduling, congestion control, resource pooling, fairness, energy saving, packet reordering, partially reliable extension, potentially failed path detection, and multipath usage. More recently, an increasing number of researchers are actively developing the promising collaborative multipath transmission technologies in order to break through the default TCP/IP stack framework and the rigid sender-centric control design.

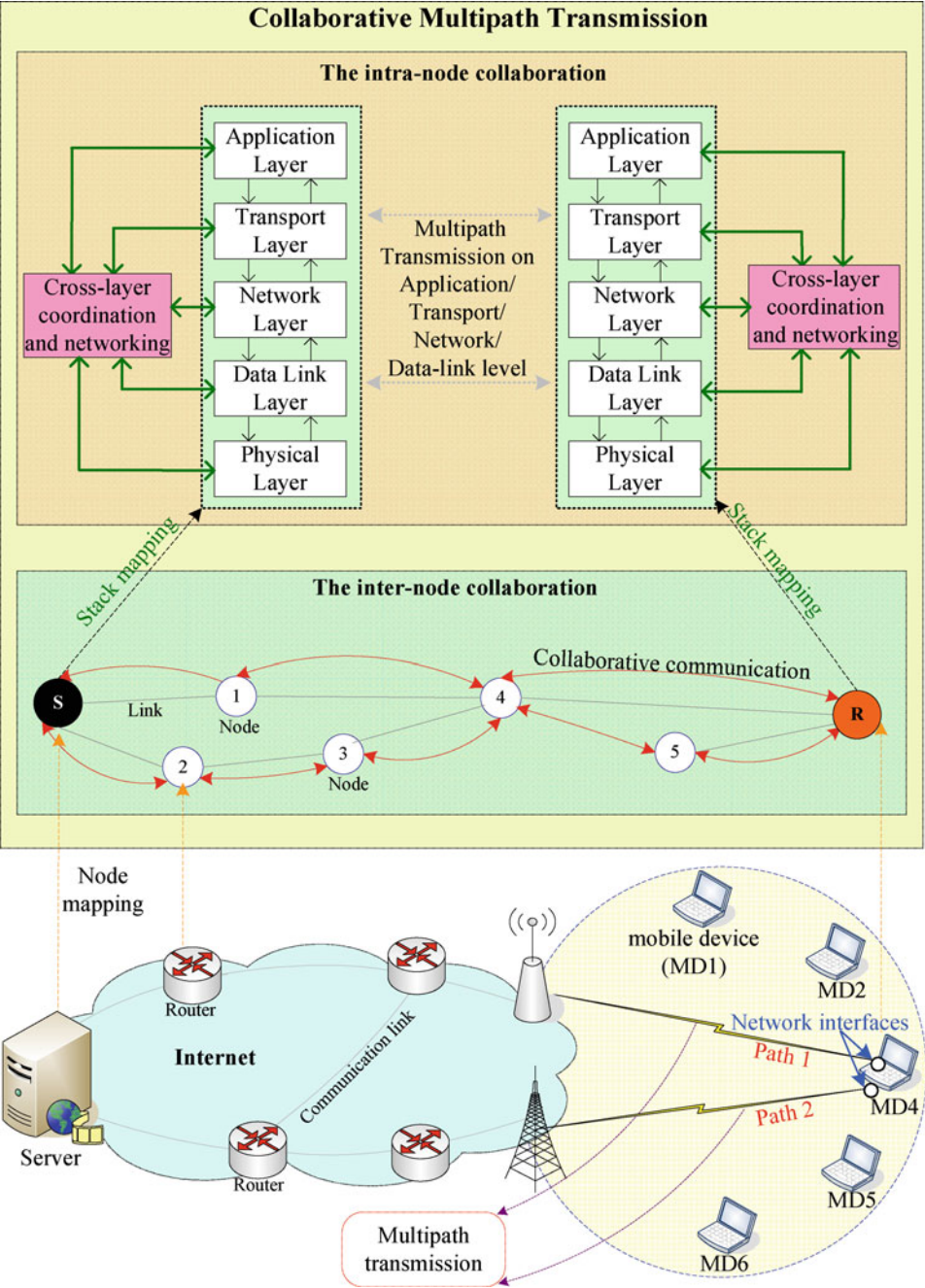
Key Points

Collaborative multipath transmission technologies can be categorized into two types: *internode collaborative multipath transmission technologies* and *intra-node collaborative multipath transmission technologies*. *Internode*

collaborative multipath transmission in general means the collaboration among various network nodes in multipath transmission (Cao et al. 2014, 2016a,b). *Intra-node collaborative multipath transmission* here refers to the cross-layer collaboration among various network protocol stacks of one node for efficient multipath transmission (Lim et al. 2014; Zorba et al. 2011; Shreedhar et al. 2018; Corbillon et al. 2016). Figure 2 presents a basic framework of collaborative multipath transmission which involves the two types of collaborative multipath transmission technologies.

Internode collaborative multipath transmission: In this situation all the network nodes within the connection session can participate in the operations of the multipath transmission. Taking the data sender and the receiver, for example, both of the two nodes can work collaboratively with each other to efficiently perform all important tasks such as path estimation and selection, flow/congestion control, and energy saving, which means, the receiver not only participates in and contributes to multipath transmission by giving feedback information to the corresponding sender in the form of acknowledgments but also acts as a decision-maker to make meaningful decision according to its obtained firsthand knowledge.

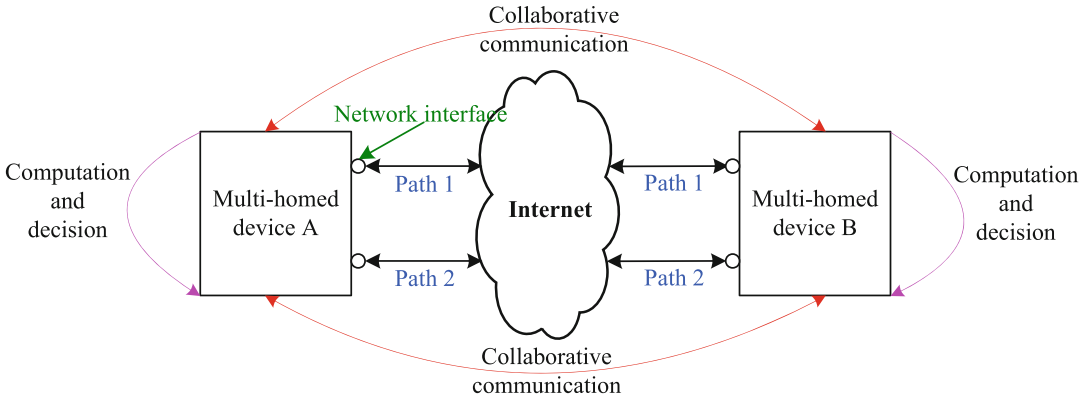
Figure 3 presents a typical scenario of internode collaborative multipath transmission between the sender and receiver sides. This collaborative



Collaborative Multipath Transmission, Fig. 2 A basic framework of collaborative multipath transmission

transmission mode has a potential for achieving many and attractive benefits (in comparison with the traditional sender-centric multipath transmission mode):

- Load balancing (Quan et al. 2017). In the traditional sender-centric mode, a sender needs to perform all important tasks and can become “bottleneck” in multipath transmission. Mov-



Collaborative Multipath Transmission, Fig. 3 Internode collaborative multipath transmission: A basic scenario

ing some tasks from the sender side onto the receiver side to make collaborative work can balance overhead and share load between the sender and receiver.

- Computing power aggregation. The collaborative multipath transmission mode not only can eliminate the performance bottleneck caused by the single sender node but also can aggregate the computing power of both the sender and receiver for effective multipath transmission.
- Proper decision-making. In practice, the receiver knows better than the sender about the characteristics of the last hop of the connection path. Thus, the receiver can perform a proper decision-making in some operations such as path quality estimation, by collaborative communicating with the sender.

Intra-node collaborative multipath transmission: As mentioned earlier, intra-node collaboration refers to cross-layer collaborative communication among each layer of the protocol stack. The benefits of applying the cross-layer collaborative transmission solution to wireless networking have been demonstrated to be very useful for achieving better performance or a global optimization of the wireless network. Compared to the traditional layered multipath transmission mode, an intra-node collaborative multipath transmission mode

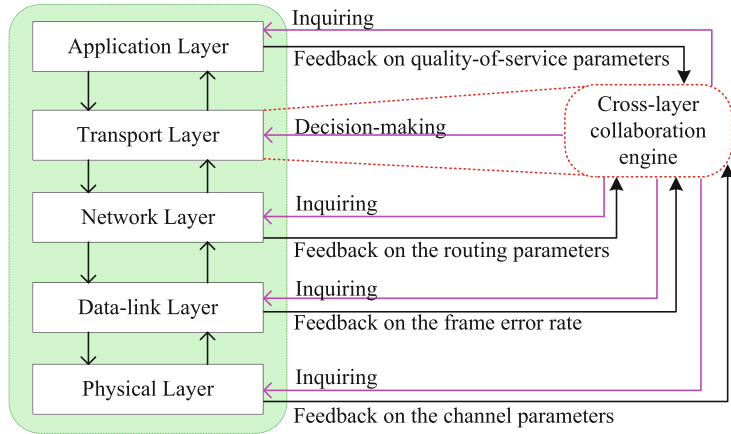
mainly has the characteristic of some following aspects:

- Message communication among various network protocol stacks does not follow strict layering principles, which means information from any one of layers in the protocol stack is contributed and made available to another layer(s).
- An efficient cross-layer adapter or engine has to be included in the networking protocol architectures to violate the conventional strict layering principles and support cross-layer collaborative communication.

We here take, for example, the transport layer-based multipath transmission protocols. Figure 4 illustrates a basic scenario of applying intra-node (cross-layer) collaboration approach to the transport layer. As the figure shows, there is a cross-layer collaboration engine that is embedded into the transport layer protocols for handling cross-layer information. By using the cross-layer collaboration engine, the transport layer can exchange information to the rest layers of the protocol stack or vice versa in order to make a proper transmission behavior or/and decision-making. For instance, the transport layer may exchange information to the application, network, data link, and physical layers in order to obtain the real-time quality-of-service parameters, routing metrics, frame error rate, and channel information,

Collaborative Multipath Transmission, Fig. 4

Intra-node (cross-layer) collaborative multipath transmission: a basic scenario



respectively, to improve the overall transmission performance.

Future Directions

The recent emergence of collaborative multipath transmission technologies has great potentials for enabling a multi-homed network device to concurrent use of multiple network links toward collaborative multipath. However, there are still several design requirements and challenges to be addressed:

- In practice, different nodes with different characteristics (i.e., computing capacity, energy power, and work load) are more commonplace. These differences among the nodes can cause performance degradations of collaborative multipath transmission. How to select a subset of suitable nodes for collaborative multipath transmission is an interesting challenge worth further investigation.
- Another concern when applying collaborative multipath transmission technologies or, more accurately, the internode collaborative multipath transmission technologies to traffic data delivery is related to the system time synchronization problem. How to make the clocks of all the nodes to be synchronized perfectly is still a hard problem.

Cross-References

- [Cognitive Heterogeneous Networks](#)
- [Network Selection in Heterogeneous Wireless Networks](#)
- [Transmission Capacity](#)

References

- 8021AX (2008) IEEE Standard for Local and Metropolitan Area Networks – Link Aggregation. IEEE Standard 8021AX
- Apple (2018) Use Multipath TCP to create backup connections for iOS. <https://support.apple.com/iv-lv/HT201373>
- Cao Y, Xu C, Guan J, Zhang H (2014) Receiver-driven SCTP-based multimedia streaming services in heterogeneous wireless networks. In: Proceedings of the IEEE international conference on multimedia and expo (IEEE ICME), pp 1–5
- Cao Y, Liu Q, Zuo Y, Ke F, Wang H, Huang M (2016a) Receiver-centric buffer blocking-aware multipath data distribution in MPTCP-based heterogeneous wireless networks. KSII Trans Internet Inf Syst 10:4642–4660
- Cao Y, Liu Q, Zuo Y, Luo G, Wang H, Huang M (2016b) Receiver-assisted cellular/wifi handover management for efficient multipath multimedia delivery in heterogeneous wireless networks. EURASIP J Wirel Commun Netw 2016:1–13
- Corbillon X, Aparicio-Pardo R, Kuhn N, Texier G, Simon G (2016) Cross-layer scheduler for video streaming over MPTCP. In: Proceedings of the ACM international conference on multimedia systems (ACM MMSys)
- Detal G, Barré S, Peirens B, Bonaventure O (2017) Leveraging multipath TCP to create hybrid access networks. In: Proceedings of the ACM SIGCOM (Industrial Demos)

- Devetak F, Kapoor S (2017) Dynamic multipath routing. IETF Internet-Draft (draft-kapoor-rtgwg-dynamic-multipath-routing-01)
- Eastlake D, Zhang M, Ghanwani A, Manral V, Banerjee A (2008) Multiconstrained QoS multipath routing in wireless sensor networks. *Wirel Netw* 14:465–478
- Eastlake D, Zhang M, Ghanwani A, Manral V, Banerjee A (2014) Transparent interconnection of lots of links (TRILL): clarifications, corrections, and updates. IETF RFC 7180
- Elgabri A, Liu K, Aggarwal V (2019) Optimized preference-aware multi-path video streaming with scalable video coding. *IEEE Trans Mobile Comput* 1–1
- Ford A, Raiciu C, Handley M, Bonaventure O (2013) TCP extensions for multipath operation with multiple addresses. IETF RFC 6824
- IEEE 802.1aq (2012) IEEE Standard for Local and Metropolitan Area Networks – Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks – Amendment 20: Shortest Path Bridging. IEEE Standard 802.1aq-2012
- Lei W, Zhang W, Liu S (2018) A framework of multipath transport system based on application-level relay (MPTS-AR). IETF Internet-Draft (draft-leiwm-tsvwg-mpts-ar-09)
- Li M, Lukyanenko A, Ou Z, Antti Y, Tarkoma S, Coudron M, Secci S (2016) Multipath transmission for the internet: a survey. *IEEE Commun Surv Tutor*
- Lim Y, Chen Y, Nahum E, Towsley D, Lee K (2014) Cross-layer path management in multi-path transport protocol for mobile devices. In: *Proceedings of the IEEE conference on computer communications (IEEE INFOCOM)*
- Quan W, Liu Y, Zhang H, Yu S (2017) Enhancing crowd collaborations for software defined vehicular networks. *IEEE Commun Mag* 55:80–86
- Samsung (2018) What is download booster feature in galaxy s5. <https://www.samsung.com/ae/support/mobile-devices/what-is-download-booster-feature-in-galaxy-s5/>
- Shreedhar T, Mohan N, Kaul S, Kangasharju J (2018) QAware: a cross-layer approach to MPTCP scheduling. In: *Proceedings of the IFIP networking conference*
- Sinky H, Hamdaoui B, Guizani M (2019) Seamless hand-offs in wireless HetNets: transport-layer challenges and multi-path TCP solutions with cross-layer awareness. *IEEE Netw* 33(2):195–201
- Stewart R (2007) Stream Control Transmission Protocol. IETF RFC 4960
- Zorba N, Skianis C, Verikoukis C (2011) Cross layer designs in WLAN systems. Troubador Publishing Ltd, Leicester

Collaborative Sensor Network Localization

- [Cooperative Localization in Wireless Sensor Networks](#)

Collusion

Fan Wu and Dan Peng
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Synonyms

[Bid rigging](#); [Cartel](#); [Collaborative cheating](#); [Price agreement](#); [Tacit collusion](#)

Definitions

In markets where game theory and mechanism design are used for resource allocation or cost sharing, agents are independent and rational, i.e., each agent aims to get higher individual utility for itself. Collusion takes place when a subset of agents deviates from the strategy of truthfully reporting their private information to manipulate the game outcomes.

Historical Background

Collusion is a pervasive problem in markets. Agents have the incentive to collude and cheat, when collusion is profitable, e.g., they can gain more group utility as well as individual utilities.

Many case studies in practice show that traditional auctions are vulnerable to collusive behaviors. For example, in previous FCC spectrum auctions (Cramton and Schwartz 2000), where the bids for spectrum licenses were large in hundreds of millions and there were no constraints on the submitted bids, bidders signaled with the last three digits of their bids and coordinated a collusive agreement on the spectrum licenses. The design of simultaneous open bidding and full information revelation makes ascending auctions vulnerable to bidder collusion.

There are also a number of research studies on collusion analysis. To facilitate peer contribution (e.g., disseminating and sharing files) and

to avoid free-riding (i.e., downloading files without uploading any in return), P2P file-sharing systems always incorporate incentive mechanisms based on virtual currency or reputation. However, measurements in Yang et al. (2005) reveal that pair-wise colluding groups exchanged data aggressively to earn uploading points. Yet, another example is in secondary spectrum markets, where second-price auctions are designed for on-demand short-term spectrum redistribution in wireless networks. While truthful auctions function well in discouraging bidders from individual cheating, they become ineffective in the case of bidder collusion. Experiments in Zhou and Zheng (2010) show that colluding groups even in small size and simple pattern can easily increase group utility and degrade auction revenue.

In wireless ad hoc networks (WANETs), source node depends on other intermediate nodes to forward data to destination node. Therefore, cooperation of wireless nodes (which belong to individual agents and perform in their own interests) is crucial for the networks to function effectively. Over the last few decades, research effort on mechanism design in this field mainly focuses on two parts: routing decision-making (Wang and Li 2006; Zhong and Wu 2010) and packet forwarding (Li and Wu 2010; Chen et al. 2011). Although truthful mechanisms guarantee that no individual agent can benefit from unilateral deviation, they are still susceptible to agent collusion. For example, if two nodes realize that they are critical nodes (i.e., their failure will divide the network to disconnected parts), they can collude to declare arbitrarily large routing costs and charge a monopoly price together.

Collusion-Resistant Mechanisms

While the colluding agents benefit from cheating, the overall system performance (e.g., efficiency, fairness, and revenue) may degrade seriously. Therefore, it is significantly important for systems to design collusion-resistant mechanisms, either making collusion difficult to form or leaving agents no incentive to collaboratively cheat.

FCC Spectrum Auction

Collusion can be mitigated by appropriately enhancing the particular rules of the auctions. In the FCC spectrum auction, the revealed information about bidding and bidder identity can affect its susceptibility to bidder collusion (Cramton and Schwartz 2000; Marshall and Marx 2009). First, by specializing the three trailing digits of its bid, a bidder in the FCC auction could tell the others its demand for some certain license and warn its rivals that they would better withdraw their bids to avoid punishment. So, cutting the trailing digits of bids, setting standardized bid increments, and limiting the number of withdraws will increase the signaling cost and thus alleviate collusion. Second, revealing bidder identities allows retaliatory bids (i.e., punishments for breaking the collusive agreement) and utility transfer (i.e., payments from the winning cartel to its cartel members). Withholding the identity information of bidders can make collusion more difficult.

Compared to the open ascending auctions, second-price auctions, where bidders submit sealed bids and the highest bidder wins and pays the second highest bid, are supposed to be more robust to collusion, since they formally avoid bid signaling and bidder identity revelation. However, most truthful mechanisms, including the celebrated Vickrey-Clarke-Groves (VCG) mechanism (Vickrey 1961; Clarke 1971; Groves 1973), are vulnerable to bidder collusion. Robinson (1984) presented the theoretical evidence.

Wireless Spectrum Allocation

To characterize the property of collusion-resistance, Goldberg and Hartline (2005) proposed a strong notion of truthfulness. A mechanism is t -truthful if no collusion group of size t or less can improve its group utility even when colluding agents can transfer utilities within the group. In this context, the only t -truthful mechanism for single-item multiunit auctions is the trivial posted price scheme. While posted price may lead to large revenue loss, mechanisms, which are t -truthful with high probability but can approximately achieve either

revenue maximization or efficiency, are more desirable.

Zhou and Zheng (2010) proposed a (t, p) -truthful mechanism to address bidder collusion in large-scale spectrum auctions. In the original pricing mechanism (Zhou et al. 2008), a winner is charged by the bid of its critical neighbor, also known as critical value (i.e., the bidder wins if bidding higher than this value and loses if bidding lower). A representative collusion pattern is Winner Critical Neighbor collusion. Winner W bids a high price to compete for the spectrum, and the critical neighbor C bids an amount, which is lower than its true value but maintains her critical position. If W wins and pays C 's bidding price, then W and C successfully gain more group utility by collusion. To mitigate collusion, the proposed mechanism divides bidders with interference into different groups, selects a price and potential winners for each group, and then selects winning groups based on their group revenue. By randomizing the winner selection and rounding the group revenue, this (t, p) -truthful mechanism makes the allocation result insensitive to bid changes and thus guarantees with a probability of at least p that no collusion group of size t or less can gain a higher group utility by bid rigging.

Yet, another work Ji and Liu (2008) proposes a multistage pricing game to address bidding rigging collusion in dynamic spectrum auctions. Buyers collude and suppress competition through bid rigging. When they get higher utilities, the allocation efficiency and seller's utility diminish. To alleviate this problem, the seller sets an optimal reverse price. Wu et al. (2009) studied loser collusion and sublease collusion. In the former case, bidders collude to raise their bids to compete for spectrum; in the latter case, bidders who win in the auction sublease their spectrum to losers to earn extra utilities. Both types of collusion will affect the efficiency, fairness, and revenue. To combat collusion, the multi-winner auction due to spectrum reusability is remodeled as a single-winner second-price auction. Bidders with negligible interference are grouped together

as virtual bidders and assigned the sum of their individual bids as virtual bids. The virtual bidder with the highest virtual bid would win the spectrum, and each bidder in this virtual group has to pay an equilibrium price. High prices make losers unlikely to aggressively raise their bids, and strict constraints on the prices leave the winners little incentive to release their spectrum. Both of the two mechanisms are immune to one or more patterns of bidder collusion.

Routing in Wireless Ad Hoc Networks

A mechanism is *group strategyproof* if no group deviation can strictly improve the utility of a single agent in the group without strictly harming another one (Jain and Mahdian 2007). In this definition, an important assumption is that the utility is not transferable. That is, an agent that benefits from collusion will not compensate another agent that suffers a loss, which is different from that of t -truthful mechanism and (t, p) -truthful mechanism.

In wireless ad hoc networks, wireless nodes rely on other intermediate nodes to forward packets. Before the forwarding is possible, source nodes have to make routing decisions about which path they prefer to take. To make routing decisions, a node needs to know the information of relaying costs of other nodes in the network. Being rational, the wireless nodes strategically maximize their own utilities and collude to gain group utility if possible. In theory, group strategyproof mechanisms are more powerful than strategyproof mechanisms in sustaining the network stability.

However, group strategyproofness cannot be achieved when utilities gained by collusion is transferrable among colluding agents (Wang and Li 2006). In other words, it is impossible to design a t -truthful pricing mechanism for the least-cost unicast routing problem in WANETs even for $t = 2$. In Zhong and Wu (2010), the assumption of transferable utility is relaxed but prevented via cryptographical techniques. When agents do not fully trust each other and cannot

convince the others of their actions, they are unlikely to transfer utilities. Even without the assumption of transferable utility, there is still no group strategyproof equilibrium.

Nevertheless, a positive result is that there exists a strong Nash equilibrium after cost discretization, and it achieves social efficiency (Zhong and Wu 2010). A *strong Nash equilibrium* is a Nash equilibrium where no group can benefit all of its members by deviation, taking the actions of its complements as given. In this context, when others tell the truth, the optimal strategy for the group is to tell the truth as well.

Packet Forwarding in Vehicular Ad Hoc Networks

While noncooperative game theory studies strategy and payoff of each individual agent, cooperative game theory considers how to enforce and sustain cooperation among agents. Typical scenarios include (but are not limited to) value allocation and cost sharing.

In cooperative games, group strategyproofness is also called coalitional rationality. In the VANET coalitional game (Chen et al. 2011), a collusion-resistant coalition that enforces all vehicles to cooperate and forward packets via opportunistic connections was proposed. In general, a coalition is a subset of vehicles, and it can obtain a value or payoff if the vehicles in the coalition cooperate with each other. By carefully designing the payoff allocation mechanism, the VANET coalitional game achieves a stable grand coalition. The grand coalition can maintain cooperation among all vehicles and thus is robust to packet forwarding collusion. Even though the payoff allocation depends on the forwarding records submitted by the individual vehicles, the mechanism avoids record cheating collusion by checking and verifying the forwarding records. Therefore, the payoff allocation mechanism achieves coalitional rationality or group strategyproofness.

Open Questions

A general question is, what other applications and scenarios in wireless networks are vulnerable to collusion? Both Nash equilibrium and dominant-strategy equilibrium are susceptible to collusion (Feigenbaum et al. 2007). Being observable or hidden, collusion is hard to detect, even in small size and simple pattern. Although several case studies have dug collusion patterns, there is still no formal definition for collusion or complete characterization of collusion patterns.

In addition, many questions about the collusion-resistant mechanisms have not been fully answered. Collusion-resistant mechanisms make collusion less attractive. However, given different objective functions and constraints, little is known about the existence of collusion-resistant mechanisms, and the necessary and sufficient conditions have not been comprehensively understood.

Cross-References

- [Auction](#)
- [Game Theory-Assisted Cooperative Spectrum Access](#)
- [Routing for Vehicular Ad Hoc Networks](#)

References

- Chen T, Zhu L, Wu F, Zhong S (2011) Stimulating cooperation in vehicular ad hoc networks: a coalitional game theoretic approach. *IEEE Trans Veh Technol* 60(2):566–579
- Clarke EH (1971) Multipart pricing of public goods. *Public Choice* 11(1):17–33
- Cramton P, Schwartz JA (2000) Collusive bidding: lessons from the FCC spectrum auctions. *Regul Econ* 17: 229–252
- Feigenbaum J, Schapira M, Shenker S (2007) Distributed Algorithmic Mechanism Design. In: *Algorithmic Game Theory*, Cambridge University Press, pp 363–384
- Goldberg AV, Hartline JD (2005) Collusion-resistant mechanisms for single-parameter agents. In: *Sixteenth*

- ACM-SIAM symposium on discrete algorithms, pp 620–629
- Groves T (1973) Incentives in teams. *Econometrica* 41(4):617–631
- Jain K, Mahdian M (2007) Cost Sharing. In: *Algorithmic Game Theory*, Cambridge University Press, pp 385–410
- Ji Z, Liu KJR (2008) Multi-stage pricing game for collusion-resistant dynamic spectrum allocation. *IEEE J Sel Areas Commun* 26(1):182–191
- Li F, Wu J (2010) FRAME: an innovative incentive scheme in vehicular networks. In: *IEEE international conference on communications*, pp 1–6
- Marshall RC, Marx LM (2009) The vulnerability of auctions to bidder collusion. *Q J Econ* 124(2):883–910
- Robinson MS (1984) Collusion and the choice of auction. *Ucla Econ Working Pap* 16(1):141–145
- Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. *J Finance* 16(1):8–37
- Wang W, Li XY (2006) Low-cost routing in selfish and rational wireless ad hoc networks. *IEEE Trans Mobile Comput* 5(5):596–607
- Wu Y, Wang B, Liu KJR, Clancy TC (2009) A scalable collusion-resistant multi-winner cognitive spectrum auction game. *IEEE Trans Commun* 57(12):3805–3816
- Yang M, Zhang Z, Li X, Dai Y (2005) An Empirical Study of Free-Riding Behavior in the Maze P2P File-Sharing System. In: *International Workshop on Peer-to-Peer Systems*, pp 182–192
- Zhong S, Wu F (2010) A collusion-resistant routing scheme for noncooperative wireless ad hoc networks. *IEEE/ACM Trans Netw* 18(2):582–595
- Zhou X, Zheng H (2010) Breaking bidder collusion in large-scale spectrum auctions. In: *Eleventh ACM international symposium on mobile ad hoc network and computing*, pp 121–130
- Zhou X, Gandhi S, Suri S, Zheng H (2008) eBay in the sky: strategy-proof wireless spectrum auctions. In: *International conference on mobile computing and networking*, pp 2–13

Combinatorial Auctions

► Budget Feasible Mechanisms

Communication-Based Train Control (CBTC)

► Interference Mitigation in Wireless Communications-Based Train Control (CBTC), Cognitive Control Approach

Communications and Networking in Droplet-Based Microfluidic Systems

Werner Haselmayr¹, Andrea Zanella², and Giacomo Morabito³

¹Johannes Kepler University Linz, Linz, Austria

²University of Padova, Padua, Italy

³University of Catania, Catania, Italy

Synonyms

[Lab-on-a-chip](#); [Tiny volumes of fluids](#); [Two-phase flow microfluidics](#)

Definition

Tiny volumes of fluids, so-called droplets, are used for communication and/or networking purposes in microfluidic chips.

Historical Background

Communications in microfluidic channels is a type of molecular communication. In Bicen and Akyildiz (2014), the transport of molecules that are dispersed into a continuous fluid in a microfluidic channel has been investigated. In droplet-based microfluidics, droplets flow in microfluidic channels inside an immiscible continuous flow. The droplets can be independently controlled and manipulated, paving the way for communications, computing, and networking in microfluidic chips.

Droplet-Based Communications

The idea of encoding/decoding of information using droplets has been, for the first time, introduced in Fuerstman et al. (2007). In particular, the distance between two droplets has been used for information encoding, and experimental results have been presented. Later on, various other encoding schemes have been defined in De Leo et al. (2012, 2013), including

presence/absence of droplets, size of droplets, and substances composing the droplets. Channel capacity expressions for presence/absence, size, and type encoding have been presented in De Leo et al. (2012, 2013), assuming no droplet coalescence. However, no closed-form expressions can be found if the aforementioned assumption is not satisfied, and, thus, a numerical analysis of the channel capacity has been presented in Galluccio et al. (2018). For information encoding using the inter-droplet distance, a closed-form expression for the channel capacity is given in De Leo et al. (2012, 2013). The error performance when information is encoded in the droplet size and the inter-droplet distance has been evaluated experimentally in Biral et al. (2015b). Moreover, the tradeoff between data rate and error probability for substance-based encoding has been investigated in Galluccio et al. (2015).

Droplet-Based Networking

The principle of controlling the droplets' path in a network of microchannels by only exploiting hydrodynamic effects has been, for the first time, introduced in Prakash and Gershenfeld (2007). In this work, this principle has been used for the implementation of logic gates. The interconnection of microfluidic devices (e.g., lab-on-a-chip (LoC)) in a microfluidic network has been proposed in De Leo et al. (2012, 2013), with the aim of realizing programmable, flexible, and biocompatible microfluidic chips. The microfluidic devices are interconnected using so-called microfluidic switches, which control the droplets' path within the network by only exploiting hydrodynamic effects. Two different types of switches have been proposed so far, namely, single-droplet switch in Cristobal et al. (2006); De Leo et al. (2013) and multiple-droplet switch in Castorina et al. (2017). The former switch controls the path of a single droplet, while the latter controls multiple droplets. The single-droplet switch has been deployed in different network architectures, using different addressing schemes to steer the droplet to the targeted microfluidic device. For example, a ring network where the distance between droplets is exploited for addressing purpose has

been reported in De Leo et al. (2012, 2013). Moreover, a bus network where the addressing is based on the droplet size and the inter-droplet distance has been introduced in Zanella and Biral (2014) and in Donvito et al. (2016), respectively. A comparison between both addressing methods has been presented in Hamidović et al. (2018).

The multiple-droplet switch can only be used in bus networks, which has been investigated in Castorina et al. (2017). In order to avoid droplet collisions in microfluidic networks, a device for medium access control has been proposed in Donvito et al. (2013). Before fabrication, the microfluidic networks are usually simulated. Unfortunately, conventional computational fluid dynamics (CFD) simulations are very time-consuming and require high computational costs. Thus, two tools especially dedicated to the simulation of microfluidic networks have been reported in Grimmer et al. (2018a) and Biral et al. (2015a), where both methods are based on the analogy between microfluidic networks and electrical circuits.

Foundations

Droplet-Based Communications

The encoding of information using droplets can be done in many different ways. In the following, a brief description of different encoding schemes is given, and the corresponding capacity/throughput formulas are presented.

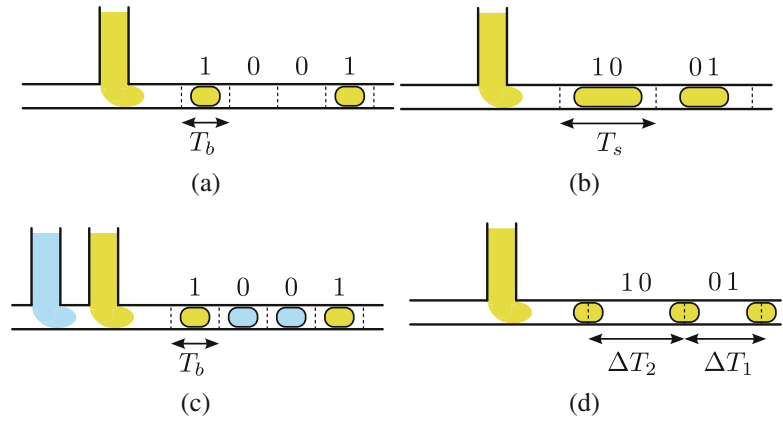
Presence/Absence of Droplets: In this case, the transmission is divided into time slots, each of duration T_b . Within this time slot, a droplet is generated for bit 1, and no droplet is generated for bit 0 (see Fig. 1a). Assuming that droplets do not leave their slot, the throughput for this encoding scheme is given in De Leo et al. (2012) as

$$C_{\text{DPA}} = \frac{1}{T_b}. \quad (1)$$

Size of Droplets: In this case, different droplet sizes are generated to convey information. Several droplet sizes can be used to encode symbols that represent several bits. For example, four

Communications and Networking in Droplet-Based Microfluidic Systems,

Fig. 1 Different droplet-based information encoding schemes for communications in microfluidic chips: (a) Presence/absence encoding; (b) size encoding; (c) composition encoding; (d) distance encoding



different droplet sizes are required to encode symbols that represent two bits (see Fig. 1b). Thus, assuming that the droplets do not leave their time slots, the throughput is given by

$$C_{DS} = \frac{m}{T_s}, \quad (2)$$

where T_s and m denote the duration of a symbol time slot and the number of bits per symbol, respectively. This encoding scheme is a generalization of the presence/absence encoding scheme.

Composition of Droplets: In this case, the information can be either encoded in different substances composing the droplet or in different molecules encapsulated inside the droplet (see Fig. 1c). Similar to size-based encoding, various substances/molecules can be used to encode symbols that represent several bits. For the binary and non-binary case, the throughput is given by (1) and (2), respectively.

It is important to note that without the assumption that droplets remain in their time slot, no closed-form expression for the capacity is available. This is because then the channel has both memory and anticipation, i.e., the output of the channel in a certain time slot depends on the droplets generated at the previous, the current, and the next slot. In Galluccio et al. (2018), an extensive analysis of this channel based on Markov chains has been presented, and the channel capacity has been evaluated numerically.

Distance Between Droplets: In this case, the information is encoded in the distance between

consecutive droplets. Several distances can be used to encode symbols that represent multiple bits. For example, to encode symbols that represent two bits, four different distances, ΔT_j , $j = 1, 2, 3, 4$, are required (see Fig. 1d). It has been shown in De Leo et al. (2013) that the channel capacity can be written as

$$C_{DD} = \frac{1}{\sqrt{2\pi e \sigma_E^2} \log 2} \exp \left(-W \left(T_{\min} / \sqrt{2\pi e \sigma_E^2} \right) \right), \quad (3)$$

where $T_{\min}$ denotes the minimum time interval between two consecutive droplets to avoid droplet coalescence, e corresponds to the Euler's number, and σ_E^2 is the variance of the random variable representing the difference between the distance at the transmitter and the one at the receiver. Moreover, $W(\cdot)$ denotes the Lambert W function, which is defined by the equation $z = W(z) \exp(W(z))$.

Droplet-Based Networking

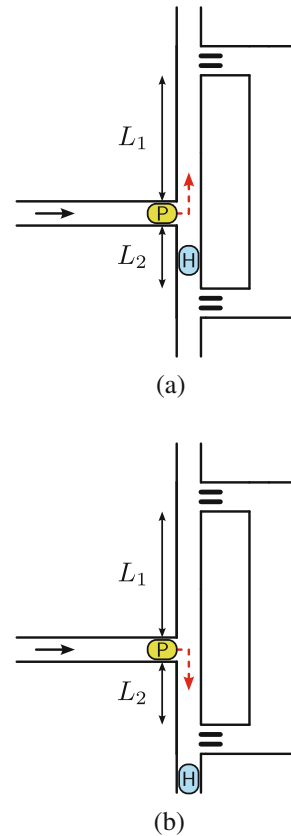
The key building blocks of droplet-based microfluidic networks are microfluidic switches. These switches control the droplet's path within a microfluidic network only exploiting hydrodynamic effects. Hence, the main design parameters are the channel geometries and hydrodynamic forces (e.g., volumetric flow rate). In the following, first, two different types of microfluidic switches are introduced, and then

their application in microfluidic bus networks is shown.

Microfluidic Switches: Microfluidic switches control the path of so-called *payload* droplets using so-called *header* droplets. Payload droplets contain chemical or biological samples that are sent to a certain network node (microfluidic device) for processing. On the other hand, header droplets contain no samples and are only used for signaling purpose. Two different types of switches have been proposed so far: single- and multiple-droplet switches, controlling the path of a single and multiple droplets, respectively.

The single-droplet switch is shown in Fig. 2 and consists of a T-junction, two asymmetric channels 1 and 2, and a bypass channel. The lengths of channels 1 and 2 are given by L_1 and L_2 , respectively. Due to the bypass channel, the switch only depends on the geometry of channels 1 and 2, which is an important property when multiple switches are cascaded in a network. It is assumed that channel 1 is longer than channel 2, i.e., the hydrodynamic resistance of channel 1 is higher than that of channel 2. If a payload droplet with a leading header droplet flows toward the T-junction, the header droplet flows into channel 2, due to its lower hydrodynamic resistance. If the following payload droplet arrives at the T-junction before the header droplet has left channel 2, it flows into channel 1 (see Fig. 2a). This is because the header droplet increased the hydrodynamic resistance of channel 2 such that it is higher than in channel 1. If the header droplet has already left channel 2, the payload droplet follows the header droplet into channel 2 (see Fig. 2b). Thus, whether the payload droplet takes channel 1 or 2 is encoded in the distance to the preceding header droplet.

The multiple-droplet switch is shown in Fig. 3 and consists of a control and a switching region. The switching region contains an asymmetric junction with channels 3 and 4, where channel 3 (length L_3) is longer than channel 4 (length L_4), i.e., channel 3 has a higher hydrodynamic resistance than channel 4. The basic idea of this switch is that a droplet in the control region controls the path of a droplet which is currently in the switching region. In particular, a droplet at the

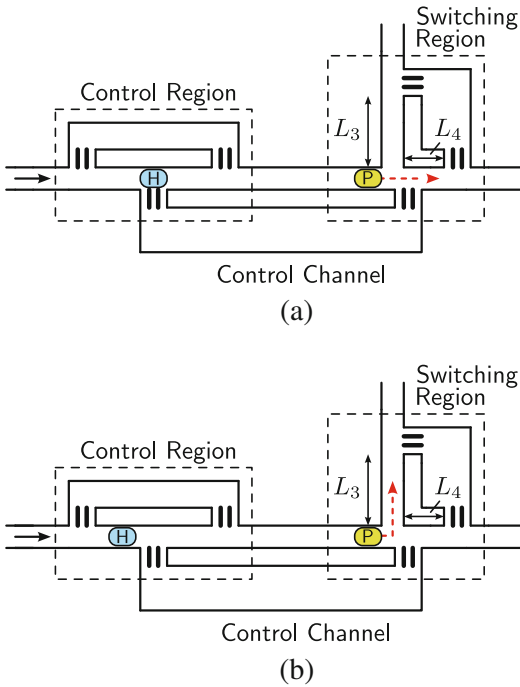


Communications and Networking in Droplet-Based Microfluidic Systems, Fig. 2 Single-droplet switch (header and payload droplets are labeled by “H” and “P”, respectively): (a) the distance between payload and header droplet before the junction is smaller than L_2 , and, thus, the droplets flow into different channels; (b) the distance between payload and header before the junction is greater than L_2 , and, thus, both droplets flow into the same channel

switching region flows into channel 4, if the following droplet clogs the control channel in the control region (see Fig. 3a). This is because the hydrodynamic resistance of channel 4 is lower than that of channel 3. On the other hand, a droplet arriving at the switching region flows into channel 3, if the control channel is not clogged by the following droplet (see Fig. 3b). This is because the flow through the control channel prevents the droplet in the switching region to flow into channel 4. Thus, similar to the single-droplet switch, the distance between the droplets determines the droplets’ path in the multiple-

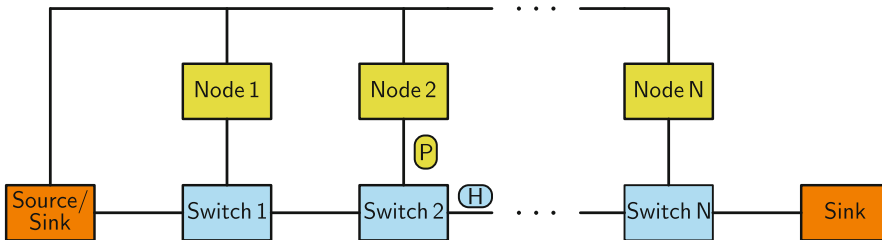
droplet switch. It is important to note that this switch can be used to control the path of a single or multiple payload droplets only using a single header droplet at the end.

Microfluidic Bus Network: Three different network topologies have been identified for microfluidic networks: bus, ring, and application-specific. In the following, a detailed discussion



Communications and Networking in Droplet-Based Microfluidic Systems, Fig. 3 Multiple-droplet switch (header and payload droplets are labeled “H” and “P”, respectively): (a) the control channel is clogged, and, thus, the payload droplet flows into channel 4; (b) the control channel is not clogged, and, thus, the payload droplet flows into channel 3

on the bus network is provided since it offers the highest flexibility and scalability compared to ring networks and application-specific architectures presented in De Leo et al. (2012) and Grimmer et al. (2018b), respectively. Moreover, only the bus network allows to apply single- and multiple-droplet switches. The topology of a microfluidic bus network is shown in Fig. 4. Considering the single-droplet switch described above, channel 1 is connected with the network node, and channel 2 is connected with the next switch in the network. The payload droplet can be routed to the targeted network node by generating a header droplet in front. The distance between both droplets determines the targeted node. In particular, if the distance between the droplets is larger than the length of channel 2, both droplets flow toward the next switch, and their distance is reduced. If their distance is below the length of channel 2, the payload droplet is routed to the desired network node for further processing, and the header droplet passes the remaining switches and flows into a sink. Considering the multiple-droplet switch described above, channel 4 is connected with the network node, and channel 3 is connected with the next switch in the network. This switch enables to route multiple payload droplets to the targeted network node with only a single header droplet at the end. This is accomplished through an equally spaced payload droplet train with a header droplet at the end. In particular, the distance between the droplets must be set, such that at the desired network node, the previous droplets clog the control channel and, thus, the droplet at the switching region flows toward the targeted node, i.e., into channel 4. It



Communications and Networking in Droplet-Based Microfluidic Systems, Fig. 4 Microfluidic bus network, where the payload droplet is routed to node 2 for further processing and the header droplet flows toward the sink

is important to note that the last payload droplet is routed by the header droplet at the end. If the control channel is not clogged by the previous droplet, the droplet train passes the switch, and the spacing between the droplets is reduced.

Key Applications

Microfluidic networking is a promising approach for applications which require a fast analysis of biochemical samples and where the viability of the sample must not be compromised. The first requirement can be achieved through parallel processing in a bus network, and the second constraint is implicitly fulfilled since the droplets' path is controlled only exploiting hydrodynamic effects. Recently, two such applications have been suggested, namely, fast and flexible drug screening in Haselmayr et al. (2018) and fast waterborne pathogen screening in Hamidović et al. (2019). The former application enables fast and flexible screening of different types/concentrations of antibiotics on infected human cells (e.g., lung cells), and the latter application proposes a fast analysis of the interaction between water pathogens and different disinfectants.

Future Directions

The research on droplet-based communications and networking in microfluidic chips is still in its infancy. In the past years, a lot of theoretical work has been devoted to this area. However, although many theoretical concepts have been introduced, an experimental validation for most concepts is still missing. An important step toward the practical realization are so-called droplet-on-demand (DoD) systems. Such systems allow the generation of droplets at prescribed times with a certain size and at a certain distance. A promising DoD system has been recently proposed in Hamidović et al. (2019), but a full characterization of this method is still missing.

The identification and practical realization of biological/chemical applications for microfluidic networks is a very important future direction. Such investigations must also include a comparison with well-established methods. In this regard, the practical realization of the applications proposed in Haselmayr et al. (2018) and Hamidović et al. (2019) would be an important step.

Recently, a promising simulation tool for microfluidic networks, based on the analogy between microfluidic networks and electrical circuits, has been presented in Grimmer et al. (2018a). It allows for fast simulation with low computational costs. However, currently only a limited number of components (e.g., droplet trapping) are supported, and, thus, the further development of this simulator is an important future direction.

Another interesting future direction would be the investigation of different network topologies (e.g., microfluidic tree network) and the development of more sophisticated scheduling algorithms. Moreover, the characterization of the network nodes (e.g., mixing) is a crucial step to enable a complete design and analysis of microfluidic networks.

Cross-References

- [Applications of Molecular Communication Systems](#)
- [Modeling Approaches for Simulating Molecular Communications](#)
- [Modulation in Molecular Signaling](#)

References

- Bicen AO, Akyildiz IF (2014) End-to-end propagation noise and memory analysis for molecular communication over microfluidic channels. *IEEE Trans Commun* 62(7):2432–2443
- Biral A, Zordan D, Zanella A (2015a) Modeling, simulation and experimentation of droplet-based microfluidic networks. *IEEE Trans Mol Biol Multi-Scale Commun* 1(2):122–134
- Biral A, Zordan D, Zanella A (2015b) Transmitting information with microfluidic systems. In: *Proceedings of the international conference on communications*, pp 1103–1108

- Castorina G, Reno M, Galluccio L, Lombardo A (2017) Microfluidic networking: switching multidroplet frames to improve signaling overhead. *Nano Commun Netw* 14:48–59
- Cristobal G, Benoit JP, Joanicot M, Ajdari A (2006) Microfluidic bypass for efficient passive regulation of droplet traffic at a junction. *Appl Phys Lett* 89(3):034104:1–034104:3
- De Leo E, Galluccio L, Lombardo A, Morabito G (2012) Networked labs-on-a-chip (NLoC): introducing networking technologies in microfluidic systems. *Nano Commun Netw* 3(4):217–228
- De Leo E, Donvito L, Galluccio L, Lombardo A, Morabito G, Zanolini LM (2013) Communications and switching in microfluidic systems: pure hydrodynamic control for networking labs-on-a-chip. *IEEE Trans Commun* 61(11):4663–4677
- Donvito L, Galluccio L, Lombardo A, Morabito G (2013) Microfluidic networks: design and simulation of pure hydrodynamic switching and medium access control. *Nano Commun Netw* 4(4):164–171
- Donvito L, Galluccio L, Lombardo A, Morabito G (2016) μ -net: a network for molecular biology applications in microfluidic chips. *IEEE/ACM Trans Netw* 24(4):2525–2538
- Fuerstman MJ, Garstecki P, Whitesides GM (2007) Coding/decoding and reversibility of droplet trains in microfluidic networks. *Science* 315(5813):828–832
- Galluccio L, Lombardo A, Morabito G, Palazzo S, Panarello C, Schembra G (2015) On the tradeoff between data rate and error probability in discrete microfluidics. In: *Proceedings of the international conference on nanoscale computing and communications*, pp 4:1–4:6
- Galluccio L, Lombardo A, Morabito G, Palazzo S, Panarello C, Schembra G (2018) Capacity of a binary droplet-based microfluidic channel with memory and anticipation for flow-induced molecular communications. *IEEE Trans Commun* 66(1):194–208
- Grimmer A, Chen X, Hamidović M, Haselmayr W, Ren CL, Wille R (2018a) Simulation before fabrication: a case study on the utilization of simulators for the design of droplet microfluidic networks. *RSC Adv* 8:34733–34742
- Grimmer A, Haselmayr W, Springer A, Wille R (2018b) Design of application-specific architectures for networked labs-on-chips. *IEEE Trans Comput-Aided Design Integr Circuits Syst* 37(1):193–02
- Hamidović M, Haselmayr W, Grimmer A, Wille R, Springer A (2018) Comparison of switching principles in microfluidic bus networks. In: *Proceedings of the international conference on nanoscale computing and communications*, pp 23:1–23:6
- Hamidović M, Haselmayr W, Grimmer A, Wille R, Springer A (2019) Passive droplet control in microfluidic networks: a survey and new perspectives on their practical realization. *Nano Commun Netw* 19:33–46
- Haselmayr W, Hamidović M, Grimmer A, Wille R (2018) Fast and flexible drug screening using a pure hydrodynamic droplet control. In: *Proceedings European conference on microfluidics*, pp 1–4
- Prakash M, Gershenfeld N (2007) Microfluidic bubble logic. *Science* 315(5813):832–835
- Zanella A, Biral A (2014) Design and analysis of a microfluidic bus network with bypass channels. In: *Proceedings of the international conference on communications*, pp 3993–3998

CoMP

- [Coordinated Multipoint Transmission and Reception](#)

Computation Offloading

- [Computation Offloading in Mobile Edge Computing](#)
- [Mining Task Offloading in Wireless Blockchain Networks](#)

Computation Offloading in Mobile Edge Computing

- Kai Peng¹, Yiwen Zhang², Xiaofei Wang³, Xiaolong Xu⁴, Xiuhua Li^{5,6}, and Victor C. M. Leung⁶
- ¹College of Engineering, Huaqiao University, Quanzhou, Fujian, P. R. China
- ²Anhui University, Hefei, Anhui, P. R. China
- ³School of Computer Science and Technology, Tianjin University, Jinnan, Tianjin, China
- ⁴Nanjing University of Information Science and Technology, Nanjing, Jiang Su, P. R. China
- ⁵Chongqing University, Chongqing, P. R. China
- ⁶The University of British Columbia, Vancouver, BC, Canada

Synonyms

[Computation offloading](#); [Mobile edge computing](#)

Definition

Mobile edge computing (MEC) is an emerging paradigm which pursues to provide better services by moving infrastructure-based cloud resources (e.g., computation, storage, bandwidth, and so on) to the edges of mobile networks. MEC is rapidly becoming a key technology of the fifth-generation (5G) mobile networks, which helps to achieve the key technical indicators of 5G business, such as ultra-low latency, ultra-high energy efficiency, and ultra-high reliability.

Computation offloading refers to that mobile users (MUs) send heavy computation tasks to edge servers (ESs) and receive the processed results from them. The goal of computation offloading is to minimize the total energy consumption or overall task execution time or both of them.

Historical Background

With the rapid development of mobile communications and the explosive usage of mobile devices carried by MUs, mobile Internet facilitates us with a pervasive and powerful platform to provide increasingly emerging applications. However, mobile devices generally have limited computation capabilities and battery power. Migrating computational applications from geographically distributed mobile devices to the infrastructure-based cloud servers has the potential to address the above issues. The remote cloud is always located in the center of core network and far away from MUs, which may cause delay fluctuation and additional transmission energy cost. In order to deal with the issue of network delay, MEC has been proposed as a new and effective paradigm (Shi et al. 2016; Wang et al. 2018a).

As illustrated in Fig. 1, the MEC architecture can be seen as a middle layer between MUs and the remote cloud which is closer to MUs and provide services for them. In addition, it can also be seen as an independent two-tier architecture consisting of MUs and ESs. Mobile devices carried

by MUs generally refer to smartphones, tablets, laptops, and so on. In addition, ESs generally refer to base stations combined with servers or cloudlets. MUs offload computing applications to ESs, which not only enables MUs to execute applications efficiently but also reduces latency caused by accessing the remote cloud. Computation offloading was originally investigated in mobile cloud computing (MCC). Computation offloading in MEC is similar to MCC, but the main difference is the offloading destination (Peng et al. 2018). It is generally assumed that the implementation of MCC computation offloading depends on a network architecture with the remote cloud, while the destination of computation offloading in MEC is ESs, such as cloudlets or base stations.

Foundations

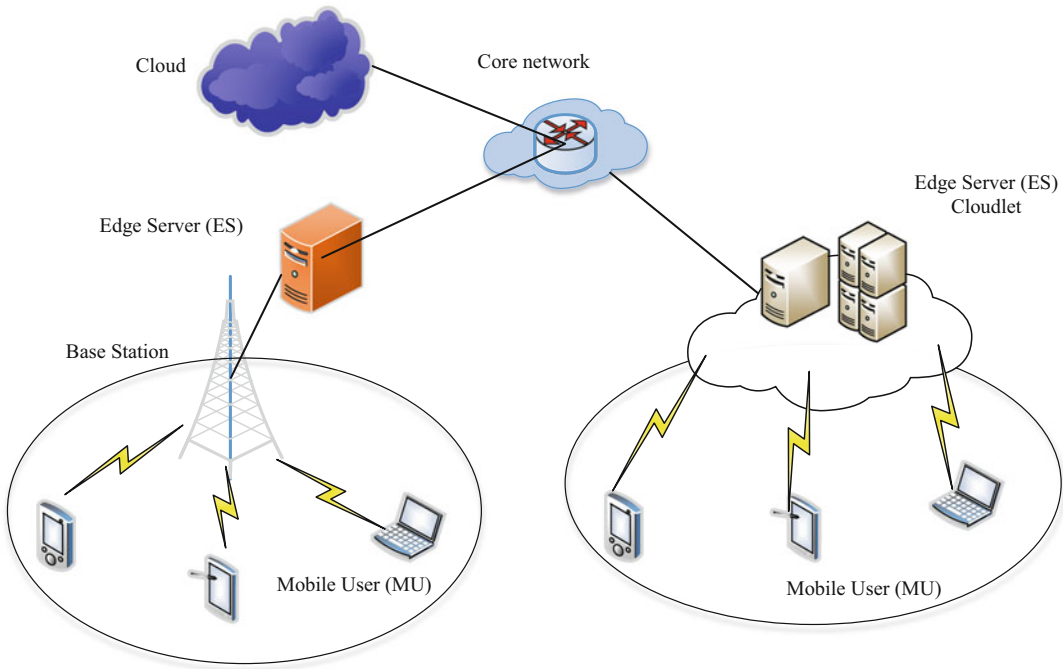
From the perspective of task division, computation offloading can be further divided into full computation offloading and partial computation offloading. From the perspective of offloading destinations (i.e., cloudlets or base stations), computation offloading can be divided into computation offloading in cloudlet-based MEC and computation offloading in base station-based MEC. More details are shown as follows.

Full Computation Offloading

Full Computation Offloading in Cloudlet-Based MEC

Full computation offloading in cloudlet-based MEC mainly focuses on the cloudlet selection among multiple cloudlets. The goal of cloudlet selection is to choose the cloudlet that meets a MU's minimum latency, or the MU's least energy consumption, or the minimum of both goals. A proxy server can be used for cloudlet selection (Mukherjee et al. 2019); however, the latency between MUs and the proxy server also needs to be considered.

If cloudlets have the same capabilities and provide the same type of services, the cloudlet selection problem may become relatively simple.



Computation Offloading in Mobile Edge Computing, Fig. 1 Mobile edge computing architecture

Namely, the nearest cloudlet is the best choice. Actually, these cloudlets are heterogeneous and have different computing capacity constraints (Xu et al. 2016), and thus it is critical to develop application-aware selection algorithm (Roy et al. 2017). More specifically, cloudlets are resource-constrained, and thus cloudlet selection and resource allocation should be taken into account jointly (Liu and Fan 2018; Zhang et al. 2018a).

VM-based cloudlet architecture is presented (Satyanarayanan et al. 2009), and a new cloudlet architecture in which applications are managed at the component level is proposed (Verbelen et al. 2012).

In addition, when the problem is switched from single-MU scenario to multi-MU scenario, the problem becomes more complicated. It is necessary to consider using game theory or deep supervised learning method to solve the problem (Yu et al. 2017; Cao and Cai 2018).

Full Computation Offloading in Base Station-Based MEC

Single-MU full computation offloading and multi-MU full computation offloading are two important issues, which need to be addressed. Different from the former mode, allocation of different types of resources, including access point association, interference management, physical resource block allocation, as well as radio resource allocation, needs to be considered in base station-based MEC.

Power consumption and the execution latency are two key goals in MEC systems (Mao et al. 2016). Moreover, minimizing the total energy consumption of access points under the MUs' individual computation latency constraints is also critical (Ketykó et al. 2016). In addition, when multiple MUs attempt to transmit data simultaneously on a shared network, the transmission collisions may occur, which is also need to be concerned (Tran and Pompili 2018).

Overall, game theory and Lyapunov optimization method as well as queuing theory can be used to figure out the above issue.

Partial Computation Offloading

Partial Computation Offloading in Cloudlet-Based MEC

Some applications, such as workflow tasks (Xu et al. 2018), can be further divided and distributed to multiple cloudlets to improve execution efficiency. In particular, some tasks can only be performed locally. Partial computation offloading can make the parallelism between mobile devices and cloudlets maximized (Jia et al. 2014).

Graph theory and 0–1 integer programming methods can be used to solve the above problem (Wu et al. 2016). Moreover, it should be noted that cloudlets are resource-constrained. In other words, tasks need to be migrated to the remote cloud or migrated back to the local execution while meeting the tasks' deadline when cloudlets cannot provide required services.

Partial Computation Offloading in Base Station-Based MEC

MUs divide tasks into sub-tasks and offload them to multiple base stations. Minimum latency and minimum energy consumption are two basic optimization goals. Meanwhile, the reliability in partial computation offloading needs to be studied (Liu and Zhang 2018). Multiple sub-parts of multiple tasks make the queue delay problem more complicated. Dynamic voltage scaling technology can be used for partial computation in base station-based MEC (Wang et al. 2016).

Key Applications

The technique of computation offloading in mobile edge computing can be utilized for improving quality of service (QoS) of MUs, such as in cloud gaming, OCR (Li et al. 2018), augmented reality (Jia and Liang 2018), and smart home (Zhang et al. 2018b), and also for

deploying future network architecture such as large-scale networks Internet of Vehicle (IoV) systems (Wang et al. 2018b).

Future Directions

From the evolution of computation in mobile edge, it appears that, in the future, emphasis will be given on the following research directions as follows:

- Deep learning-based computation optimization
- Joint computation offloading and resource allocation
- Cross layer resource optimization
- Multi-cloudlet resource optimization

Cross-References

- [Edge Computing](#)
- [Fog Computing](#)
- [Mobile Cloud Computing](#)

References

- Cao H, Cai J (2018) Distributed multiuser computation offloading for cloudlet-based mobile cloud computing: a game-theoretic machine learning approach. *IEEE Trans Veh Technol* 67(1):752–764
- Jia M, Liang W (2018) Delay-sensitive multiplayer augmented reality game planning in mobile edge computing. In: *Proceedings of the 21st ACM international conference on modeling, analysis and simulation of wireless and mobile systems*, Oct 2018. ACM, pp 147–154
- Jia M, Cao J, Yang L (2014). Heuristic offloading of concurrent tasks for computation-intensive applications in mobile cloud computing. In: *2014 IEEE conference on computer communications workshops (INFOCOM WKSHPs)*, Apr 2014. IEEE, pp 352–357
- Ketykó I, Kecskés L, Nemes C, Farkas L (2016) Multi-user computation offloading as multiple knapsack problem for 5G mobile edge computing. In: *2016 European conference on networks and communications (EuCNC)*, June 2016. IEEE, pp 225–229

- Li B, He M, Wu W, Sangaiah AK, Jeon G (2018) Computation offloading algorithm for arbitrarily divisible applications in mobile edge computing environments: an OCR case. *Sustainability* 10(5):1611
- Liu L, Fan Q (2018) Resource allocation optimization based on mixed integer linear programming in the multi-cloudlet environment. *IEEE Access* 6:24533–24542
- Liu J, Zhang Q (2018) Offloading schemes in mobile edge computing for ultra-reliable low latency communications. *IEEE Access* 6:12825–12837
- Mao Y, Zhang J, Song SH, Letaief KB (2016) Power-delay tradeoff in multi-user mobile-edge computing systems. In: 2016 IEEE global communications conference (GLOBECOM), Dec 2016. IEEE, pp 1–6
- Mukherjee A, De D, Roy DG (2019) A power and latency aware cloudlet selection strategy for multi-cloudlet environment. *IEEE Trans Cloud Comput* 7(1):141–154
- Peng K, Leung V, Xu X, Zheng L, Wang J, Huang Q (2018) A survey on mobile edge computing: focusing on service adoption and provision. *Wirel Commun Mob Comput* 2018:1–16
- Roy DG, De D, Mukherjee A, Buyya R (2017) Application-aware cloudlet selection for computation offloading in multi-cloudlet environment. *J Supercomput* 73(4):1672–1690
- Satyanarayanan M, Bahl V, Caceres R, Davies N (2009) The case for VM-based cloudlets in mobile computing. *IEEE Pervasive Comput* 8:14–23
- Shi W, Cao J, Zhang Q, Li Y, Xu L (2016) Edge computing: vision and challenges. *IEEE Internet Things J* 3(5):637–646
- Tran TX, Pompili D (2018) Joint task offloading and resource allocation for multi-server mobile-edge computing networks. *IEEE Trans Veh Technol*
- Verbelen T, Simoens P, De Turck F, Dhoedt B (2012) Cloudlets: bringing the cloud to the mobile user. In: Proceedings of the third ACM workshop on mobile cloud computing and services, June 2012. ACM, pp 29–36
- Wang Y, Sheng M, Wang X, Wang L, Li J (2016) Mobile-edge computing: partial computation offloading using dynamic voltage scaling. *IEEE Trans Commun* 64(10):4268–4282
- Wang S, Chou W, Wong KS, Zhou A, Leung VC (2018a) Service migration in mobile edge computing. *Wirel Commun Mob Comput* 2018:1–2
- Wang K, Yin H, Quan W, Min G (2018b) Enabling collaborative edge computing for software defined vehicular networks. *IEEE Netw* 32:112–117
- Wu H, Knottenbelt W, Wolter K, Sun Y (2016) An optimal offloading partitioning algorithm in mobile cloud computing. In: International conference on quantitative evaluation of systems, Aug 2016. Springer, Cham, pp 311–328
- Xu Z, Liang W, Xu W, Jia M, Guo S (2016) Efficient algorithms for capacitated cloudlet placements. *IEEE Trans Parallel Distrib Syst* 27(10):2866–2880
- Xu X, Fu S, Yuan Y, Luo Y, Qi L, Lin W, Dou W (2018) Multiobjective computation offloading for workflow management in cloudlet-based mobile cloud using NSGA-II. *Comput Intell* 35:476–495
- Yu S, Wang X, Langar R (2017) Computation offloading for mobile edge computing: a deep learning approach. In: 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), Oct 2017. IEEE, pp 1–6
- Zhang Y, Wang K, Zhou Y, He Q (2018a) Enhanced adaptive cloudlet placement approach for mobile application on spark. *Secur Commun Netw* 2018:1–12
- Zhang J, Zhou Z, Li S, Gan L, Zhang X, Qi L, Xu X, Dou W (2018b) Hybrid computation offloading for smart home automation in mobile cloud computing. *Pers Ubiquit Comput* 22(1):121–134

Computational Intelligence

- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)

Configuration and Management

- [Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization](#)

Connectedness of Molecular Nanonetworks

- [Connectivity via Molecular Signaling](#)

Connectivity via Molecular Signaling

Dogu Arifler¹ and Dizem Arifler²

¹Eastern Mediterranean University, Northern Cyprus, Turkey

²Middle East Technical University, Northern Cyprus Campus, Turkey

Synonyms

Connectedness of Molecular Nanonetworks

Definitions

Connectivity refers to a global structural network property that enables exchange of information between every pair of nodes in the network. In a nanonetwork, the nodes or nanomachines making up the nanonetwork can employ molecular signaling as a means of communication. In the most general diffusion-based molecular signaling, the transmitter nanomachine releases information-carrying molecules that diffuse in a medium to reach a receiver nanomachine. If molecular signals cannot be detected by a destination nanomachine directly, intermediate nanomachines can relay the information toward the destination also by emitting molecular signals. Therefore, if exchange of information between every pair of nanomachines is possible, either directly or indirectly, through the release and reception of molecules, the molecular nanonetwork is said to be connected.

Historical Background

Recent developments in nanotechnology enable the design and manufacture of nanomachines or devices that are made up of nanoscale components. By themselves, nanomachines can perform only simple computation, sensing, and actuation.

An interconnection of nanomachines is usually formed to enable cooperation and sharing of information to achieve more complex and useful tasks. Such nanonetworks have a wide range of potential applications in biomedicine, industry, environment, and military (Akyildiz et al. 2011).

Two main alternative communication paradigms considered for nanonetworks are nano-electromagnetic communication and molecular communication (Akyildiz et al. 2011). A feasible and biocompatible means of information transport in nanonetworks with no infrastructure can possibly be free diffusion-based molecular signaling. It is conjectured that biologically inspired nanomachines with molecular communication capabilities can be built by genetic modification of existing cells or by manufacturing artificial cells. One can refer to Farsad et al. (2016) for a very extensive survey on recent results and issues on molecular nanonetworks.

Molecular signaling mechanisms can especially be useful for broadcasting an alarm by transmitting a “puff” of molecules that will freely diffuse in a fluidic medium. Such systems can be employed for detection and handling of abnormal events, pathogens, and cancerous cells in some part of the body. Molecular concentration detectors built into nanomachines can decide whether the signal reception is successful or not based on a prespecified detection threshold. Upon detection of the molecular signal, receiver nanomachines can transmit their own molecules into the fluidic medium in a coordinated manner to percolate the alarm throughout the network provided that the network is connected. Hence, the conditions under which connectivity is achieved must be thoroughly investigated for these types of nanonetworks.

Connectivity of ad hoc wireless and sensor networks is a well-investigated subject; for example, see Krishnamachari (2005), Franceschetti and Meester (2008), and Haenggi (2012), and the references therein. However, there is relatively less work on connectivity properties of molecular nanonetworks. A recent study by Arifler (2017)

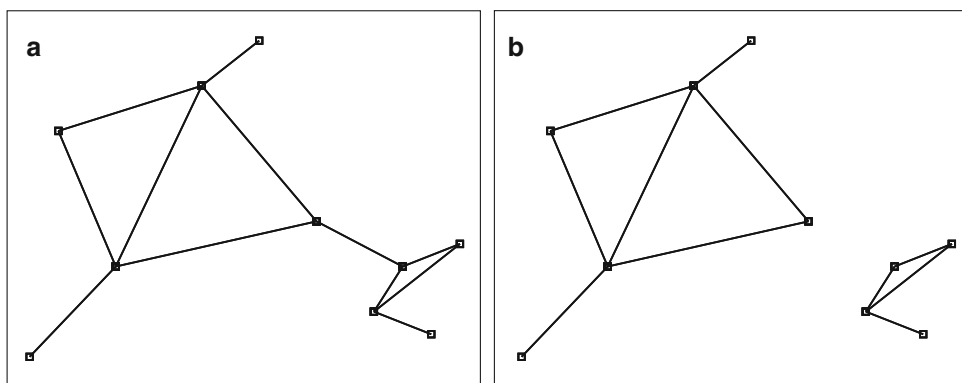
identifies differences between ad hoc wireless network and molecular nanonetwork connectivity and analyzes the connectivity properties of nanonetworks that employ free diffusion-based molecular communication.

Foundations

A network has the highest utility when any two nodes can communicate with each other. Connectivity is a graph theoretic concept (Rosen 2012). A network can be modeled as an undirected graph where the vertices are the network nodes and there is an edge between a pair of vertices if the corresponding nodes can directly communicate with each other; in other words, the edges have no orientation and correspond to communication links. A path is a sequence of edges that starts at a vertex and follows vertices along these edges of the graph. Two network nodes are connected if there is a path between them. A network is said to be connected if all pairs of nodes in the network are connected. Figure 1 depicts a connected network (Fig. 1a) and a disconnected network (Fig. 1b). Note that having all pairs of nodes connected is a strict requirement, and in practice, an almost-connected network (e.g., having at least 99% of the nodes connected) may be of greater interest.

Ad hoc wireless and sensor networks are generally modeled as two-dimensional planar networks, and the signal strength at a receiver is

inversely proportional to the α th power of the distance from the transmitting source, where α is the path loss exponent and is generally greater than 2. The nodes in large-scale networks are usually scattered randomly into an environment. Hence, a homogeneous Poisson point process can be used to model locations of randomly deployed nodes. With node locations as the vertices and direct links as the edges, the graph so formed is a random geometric graph (Haenggi 2012). In widely studied random graph models of wireless networks, there are critical values for the transmission power and the number of nodes above which the network is connected and below which the network is disconnected with high probability. Such a phenomenon of an abrupt change in the global behavior of a system in response to a small change in one of its parameters is called a phase transition (Franceschetti and Meester 2008). In a nanonetwork, the locations of nanomachines scattered randomly into a medium can also be modeled by a homogeneous Poisson point process. Yet, the nature of signal propagation in free diffusion-based molecular communication is very different: the strength of the signal observed at a receiver nanomachine is inversely proportional to the d th power of the distance from the transmitter, where d is the dimensionality of the medium. Hence, an investigation of the phase transition behavior specific to nanonetworks is necessary. In particular, determination of critical values of important nanonetwork parameters is of great interest since a nanonetwork might



Connectivity via Molecular Signaling, Fig. 1 (a) A connected network where there is a path between any two nodes and (b) a disconnected network

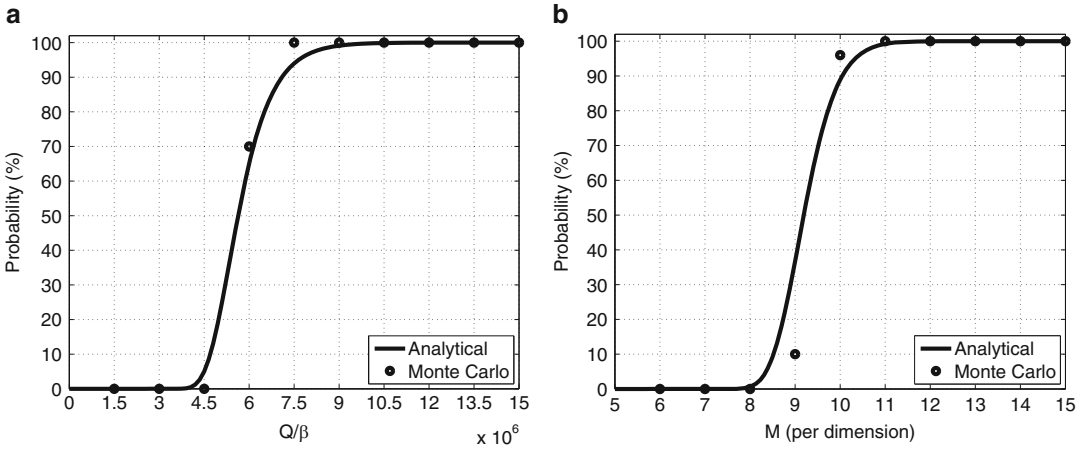
be disconnected and hence be of limited use if under-provisioned, or a nanonetwork might consume more resources than necessary if over-provisioned.

Perfectly monitoring spheres (Endres 2013) can model nanomachines that continuously and passively count molecules within their transparent volumes without disrupting the diffusion process. This type of receiver nanomachines can correspond to chemosensory detection systems in which ligands freely enter and leave the perireceptor space in a fast and reversible manner (Rospars et al. 2000). It may be assumed that nanomachines can instantly count the number of molecules within their volumes to estimate the concentration at their centers, but the molecule count is inherently noisy due to diffusion. Analogous to traditional communication systems, one can define the instantaneous signal-to-noise ratio (SNR) at a receiver nanomachine as the “power” of the molecular signal component divided by the power (variance) of the noise component in the received molecular concentration.

When communicating, each nanomachine can release particles with its self-identifying label that can be distinguished by receivers. Such a medium access mechanism has been termed as molecular division multiple access (MDMA) and was discussed in Giné and Akyildiz (2009) and Kuscu and Akan (2016). In a single “puff,” each nanomachine emits Q particles (also referred to as the signal strength), and it can be assumed that all the nanomachines in the network have the same SNR detection threshold β . It can also be assumed that a previously emitted “puff” from a given nanomachine fades out before another one can be emitted by that nanomachine. Within this framework, the SNR at a receiver centered at $\mathbf{r} \in \mathbb{R}^d$ relative to the transmitter can be determined. With the assumption of self-identifying labels, one can say that nodes that are separated by a distance $\|\mathbf{r}\|$ are directly connected if the maximum SNR detected is greater than β . Monte Carlo simulation results in Arifler (2017) indicate that a phase transition behavior similar to that in ad hoc wireless networks is also observable in

nanonetworks as the parameters signal strength-to-detection threshold ratio Q/β and the number N of randomly deployed nanomachines are varied. For illustrative purposes, Fig. 2 shows connectivity probabilities for three-dimensional nanonetwork scenarios with prespecified parameters. Note that, as shown also in Fig. 2, these results are consistent with the probabilities of having a connected network obtained via analytical models based on the diffusion equation in conjunction with the assumption of a Poisson point process as applied to infinitely large networks. Further details and the results for one- and two-dimensional nanonetworks can be found in Arifler (2017).

The time required to broadcast an alarm in a network is another important performance indicator. A simulation-based analysis can again be used to quantify this indicator for molecular nanonetworks with perfectly monitoring receivers. In a recently investigated scenario (Arifler 2017), the originator of the alarm is located at the geometric center of a given region, and the rest of the nanomachines are deployed completely randomly. The time scale at which consecutive alarm pulses are transmitted by the originator can be assumed to be much larger than the time required for the pulses to fade out in the medium. The nanomachines transmit their own signals (only once) as soon as the SNR detected exceeds a prespecified threshold. In this particular scenario, all the nanomachines are assumed to transmit the same type of molecule. At a given time instant, the concentration at a receiver due to activated transmitters is merely a superposition of all their contributions. Hence, multiuser interference occurs, but this interference is *constructive*; emissions from multiple nanomachines enhance the signal at a receiver and help detect the alarm at an earlier time. The results show that the broadcast time decreases with increasing number of nanomachines in the network. Power-law relationships between the time to broadcast an alarm and the number of nanomachines deployed can be postulated for fixed regions with different dimensionalities. In addition, for a given total molecule budget, having a



Connectivity via Molecular Signaling, Fig. 2 Probability of having an almost-connected network as a function of (a) signal strength-to-detection threshold ratio Q/β with $N = 1000$ nanomachines and (b) M ($N = M^3$) with $Q/\beta = 7 \times 10^6$ in a three-dimensional ($1 \text{ mm} \times 1 \text{ mm} \times 1 \text{ mm}$) region where the diffusion coefficient is $10^{-9} \text{ m}^2/\text{s}$ and spherical receiver nanomachines have a radius of $1 \mu\text{m}$. Note that the

Monte Carlo simulation results shown are based on 50 independent runs, each with a different realization of nanomachine positions; the probabilities of an almost-connected network are estimated as the fraction of realizations with at least 99% of the nodes connected. The probabilities of having a connected network obtained via analytical models are also included for comparison. (Adapted from Arifler 2017)

network is extremely beneficial for speeding up diffusion-based propagation of information over a deployment region.

Key Points

The dependence of network connectivity on signal strength and detection threshold has different functional forms for ad hoc wireless networks and molecular nanonetworks. Large-scale two- and three-dimensional networks modeled as random geometric graphs are connected with high probability when the probability of having an isolated node goes to zero (Penrose 1999). With an efficient medium access mechanism in effect, the exponential decay rate of the probability of having an isolated node as a function of signal strength-to-detection threshold ratio is sublinear for ad hoc wireless networks but is linear for molecular nanonetworks (Arifler 2017). This suggests that connectivity can be achieved relatively *more easily* in nanonetworks with increasing signal strength-to-detection threshold ratio.

In ad hoc wireless networks, multiuser interference is detrimental for communication, and interference mitigation techniques are often necessary. In molecular nanonetworks, multiuser interference may prove to be *beneficial* for particular application scenarios that are considered for nanonetworks, such as for propagating an alarm signal throughout the network. Indeed, the empirical studies outlined here show that interference *improves* connectivity in these types of nanonetworks. One may also refer to Deng et al. (2017) for a stochastic geometry-based analysis of superposition of contributions at a single receiver from many molecular transmitters, all of which are transmitting the same signal at the same time.

Due to analytical tractability, Arifler (2017) considers connectivity of nanomachines that perfectly monitor diffusing molecules. An alternative scenario would involve the use of receiver nanomachines that absorb molecules. However, it is conjectured that analytical modeling with many absorbing receivers is nontrivial due to spatial and temporal interdependence among receivers. Monte Carlo simulation schemes

that allow for analysis of molecule absorption probabilities in nanonetworks with multiple absorbing receivers (Arifler and Arifler 2017) can be potentially employed to analyze connectivity in such scenarios.

In summary, due to the inherent phase transition phenomenon in random networks, careful design considerations are required to ensure that a molecular nanonetwork is connected. In particular, it is very important to be able to predict the critical values of signal strength-to-detection threshold ratio and the number of nanomachines to be deployed for which nanonetwork connectivity is achieved. The studies outlined here can thus provide useful design guidelines for nanonetwork deployment.

Cross-References

- [Applications of Molecular Communication Systems](#)
- [Brownian Motion](#)
- [Nanonetworks](#)

References

- Akyildiz IF, Jornet JM, Pierobon M (2011) Nanonetworks: a new frontier in communications. *Commun ACM* 54(11):84–89
- Arifler D (2017) Connectivity properties of free diffusion-based molecular nanoscale communication networks. *IEEE Trans Commun* 65(4):1686–1695
- Arifler D, Arifler D (2017) Monte Carlo analysis of molecule absorption probabilities in diffusion-based nanoscale communication systems with multiple receivers. *IEEE Trans Nanobioscience* 16(3):157–165
- Deng Y, Noel A, Guo W, Nallanathan A, El Kashlan M (2017) Analyzing large-scale multiuser molecular communication via 3-D stochastic geometry. *IEEE Trans Mol Biol Multi-Scale Commun* 3(2):118–133
- Endres RG (2013) *Physical principles in sensing and signaling: with an introduction to modeling in biology*. Oxford University Press, Oxford
- Farsad N, Yilmaz HB, Eckford A, Chae CB, Guo W (2016) A comprehensive survey of recent advancements in molecular communication. *IEEE Commun Surv Tutorials* 18(3):1887–1919
- Franceschetti M, Meester R (2008) *Random networks for communication: from statistical physics to information systems*. Cambridge University Press, Cambridge
- Giné LP, Akyildiz IF (2009) Molecular communication options for long range nanonetworks. *Comput Netw* 53(16):2753–2766
- Haenggi M (2012) *Stochastic geometry for wireless networks*. Cambridge University Press, Cambridge
- Krishnamachari B (2005) *Networking wireless sensors*. Cambridge University Press, Cambridge
- Kuscu M, Akan OB (2016) On the physical design of molecular communication receiver based on nanoscale biosensors. *IEEE Sensors J* 16(8):2228–2243
- Penrose MD (1999) On k -connectivity for a geometric random graph. *Random Struct Algorithms* 15(2):145–164
- Rosen K (2012) *Discrete mathematics and its applications*, 7th edn. McGraw-Hill, New York
- Rospars JP, Krivan V, Lánský P (2000) Perireceptor and receptor events in olfaction. Comparison of concentration and flux detectors: a modeling study. *Chem Senses* 25:293–311

Consistency, Integrity

- [Wireless Sensor Network Security](#)

Contact Plan Design for Space Disruption/Delay-Tolerant Networks

Shuang Xu and Xingwei Wang
College of Computer Science and Engineering,
Northeastern University, Shenyang, China

Synonyms

[Contact plan determination](#); [Link assignment](#); [Topology control](#); [Topology design](#)

Definitions

Contact Plan Design (CPD) refers to select the most appropriate set of feasible communication

This work is supported by the National Natural Science Foundation of China under Grant No. 61572123; the Program for Liaoning Innovative Research Term in University under Grant No. LT2016007.

opportunities in a time-evolving and predictable Delay/Disruption Tolerant Networks (DTN) to deliver data, with fulfilling certain criteria subject to restrictions or limitations.

Historical Background

Traditional geostationary Earth orbit (GEO) satellite relay systems suffer from long round-trip time and frequent channel disruption when they are used for bidirectional and interactive data communication. In deep space systems, longer distance creates higher delay, and planet rotation causes more server disruptions; thus, it is impossible to achieve permanent and conversational communications. Even in low Earth orbit (LEO) satellite systems, disruptions are prevalent, as the high complexity and huge investment are required to maintain end-to-end connectivity. As a result, TCP/IP-based Internet protocols are ineffective against these challenges in GEO, LEO, or deep space systems.

The DTN architecture, which overlays a bundle layer between application layer and transport layer, is recognized as a promising solution for future satellite networks (Caini et al. 2011). It overcomes the delay/disruption problems by using store-carry-forward message scheme. The effectiveness of the DTN paradigm over operational space networks has been tested in real experiments, such as DINET experience and JAXA DRTS testing.

Network topology is a key operational issue in space network design. For different space missions, network topology is controlled under different objectives. Considering the orbital dynamics of space DTN nodes, the contacts (i.e., communication opportunity) among space nodes only can be established when the physical requirements (such as visibility, received power, and antenna pointing) are satisfied. Besides, due to the resource constraints in satellites (such as available transponders and power consumption), not all feasible contacts can be used to route data (Fraire et al. 2013). What's more, different selection criteria (such as capacity and fairness) need to be considered to select the scheduled

contacts for different space applications (Fraire and Finochietto 2015b). This problem is defined as contact plan design (CPD). A short list of representative works on CPD is presented below.

Huang et al. (2010, 2013) proved that the time-evolving graph-based CPD in time-evolving and predictable DTN networks was an NP-hard problem and proposed a greedy-based method which minimized path costs and maximized network connectivity. Based on the min-max fairness, Fraire et al. (2013, 2014) designed a fair-aware CPD strategy that can maximize the global system fairness and the overall capacity simultaneously. However, these two studies assumed that the bandwidth of communication link was constant and the traffic distribution among nodes was uniform.

Wang et al. (2016) considered the time-varying capacities of inter-satellite and satellite-ground contacts to formulate a long-term network throughput maximization problem and derived the optimal contact plan by using efficient heuristic algorithms. Zhou et al. (2017) designed a mission-aware contact plan strategy for small satellite networks, which simultaneously considered mission characteristics, time-varying contact capacity, resource constraints, and basic constraints.

Fraire and Finochietto (2015a) incorporated route information into CPD to design a routing-aware fair contact plan strategy optimizing the fairness and route delays. The traffic of Earth observation mission is predictable, so Fraire et al. (2016) exploited the predictable traffic information to design a traffic-aware contact plan (TACP) strategy, enhancing the schemes proposed in Fraire and Finochietto (2015a).

Contact plans are commonly pre-calculated by the ground control center and delivered to orbiting satellites timely. This implies that valuable and efficient contact plans need to be found out in a reasonable time. To this end, J.A. Fraire et al. proposed a heuristic approach based on evolutionary algorithms to provide sub-optimal yet efficient contact plans in a reasonable time (Fraire et al. 2017, 2015), extending the applicability of TACP to support real and complex satellite DTN missions operation.

Foundations

Terminology

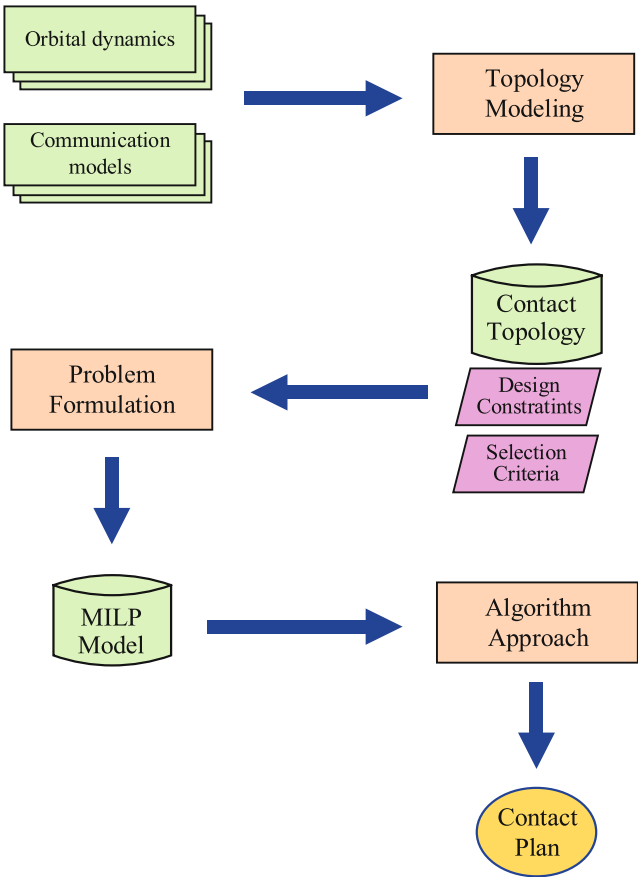
- **Contact:** The contact is a temporal communication opportunity between two DTN nodes. At least a start and stop time and a transmitting and receiving node pair are required to describe a contact. These parameters are determined by orbital dynamics (such as position, range, and attitude) and communication models (such as bit error rate, transmission power, and modulation).
- **Contact topology:** In a given DTN network, the contact topology is formed by all the feasible contacts within a given time interval. In the contact topology, no resource limitation is assumed on nodes, and just the feasibility of physical communication is considered.

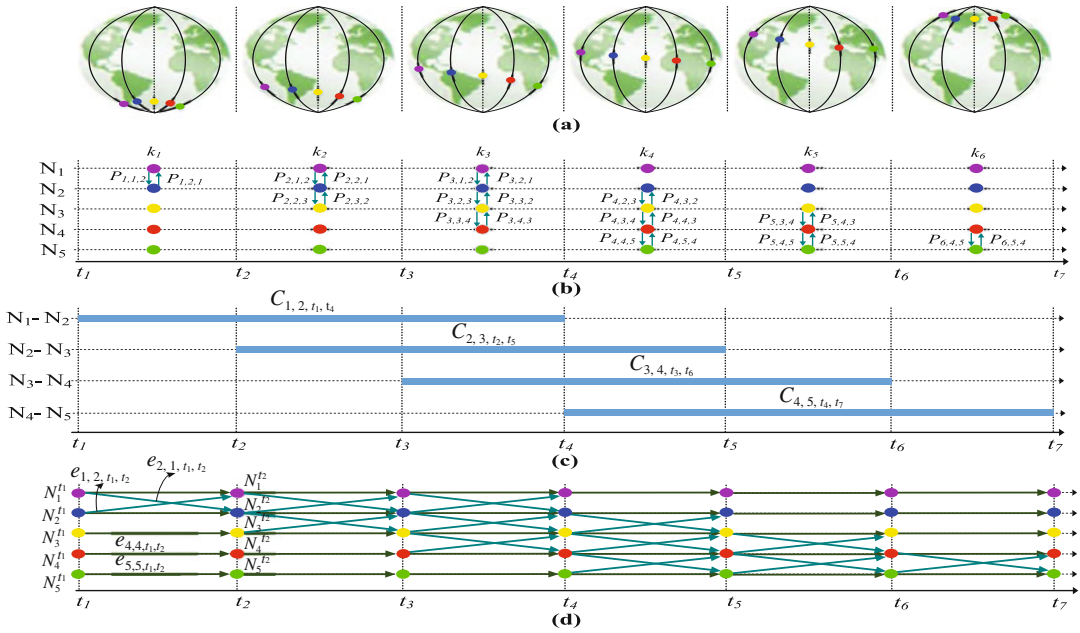
- **Contact plan:** Contact plan comprises all forthcoming contacts which satisfy the constraints and special criteria. It is a subset of the contact topology.

Contact Plan Design

In general, the process of CPD can be divided into three stages: topology modeling, problem formulation, and problem-solving. In the initial stage, based on the orbital dynamics and communication subsystem attributes, the feasibility of all communication opportunities is determined to form a contact topology. Then, the contact topology, resource restrictions, and special criteria are taken to formulate the problem model. Finally, algorithm solutions with low computation complexity are designed to solve the problem model to deliver an efficient and implementable contact plan. The process of CPD is illustrated in Fig. 1.

Contact Plan Design for Space Disruption/Delay-Tolerant Networks, Fig. 1
The processing steps of CPD





Contact Plan Design for Space Disruption/Delay-Tolerant Networks, Fig. 2 Topology model of an example satellite networks: (a) five-satellite network, (b) finite

state machine model, (c) contact list model, and (d) space-time model

Contact Topology Modeling Technique

Topology modeling is fundamental to characterize the time-evolving and predictable space DTN networks. A five-satellite network is taken as an example to explain the popular modeling methods, and the corresponding time-evolving topologies are shown in Fig. 2a.

Finite state machine The time-evolving topology shown in Fig. 2a can be characterized by a sequence of graphs, whose vertices and edges, respectively, represent the DTN nodes and feasible contacts among DTN nodes. Specially, the dynamic topology is discretized into K states separated by K time intervals. Each state can be identified by a graph which represents the static topology over a given interval. The feasibility of link between node N_i and N_j at state k can be characterized by a variable $p_{k,i,j}$. As a result, the contact topology can be defined by a three-dimensional adjacency matrix. This representation can be thought as a finite state machine model, as illustrated in Fig. 2b.

Contact list The time-evolving topology can also be characterized by a contact list where each

contact is defined by the starting and end time and the transmitting and receiving nodes. The contact between node N_i and N_j over the time interval $[t_k, t_{k+1}]$ is indicated by $C_{i,j,t_k,t_{k+1}}$. The contact list modeling for the example topology can be represented by four contacts, as illustrated in Fig. 2c.

Space-time graph Alternatively, space-time graph, a directed graph defined in both spatial and temporal space, can be exploited to characterize the evolving topology and network resources. The directed graph is composed of $K + 1$ layers, and each layer corresponds to a network topology at a given time point. The i -th node at time t_k can be denoted by $N_i^{t_k}$. A temporal edge $e_{i,i,t_k,t_{k+1}}$ represents whether node $N_i^{t_k}$ can store and carry data in the k -th time interval $[t_k, t_{k+1}]$. A spatial edge $e_{i,j,t_k,t_{k+1}}$ represents whether node $N_i^{t_k}$ can forward data to node $N_j^{t_{k+1}}$ during the time interval $[t_k, t_{k+1}]$. The space-time graph model of the example topology is illustrated in Fig. 2d.

The contact list model can be expressed by a contact table (Araniti et al. 2015); thus, it is more

compact than finite state machine and space-time graph. However, more detailed and precise topology description can be provided by finite state machine and space-time graph. Thus, finite state machine and space-time graph are convenient for CPD with using MILP optimization. Nonetheless, no matter which modeling technique is taken, the transition among them is straightforward.

Design Constraints and Selection Criteria

Except the contact topology, design constraints and special criteria are generally considered during the CPD problem formulation. The design constraints can be categorized into two groups. One is to determine the feasibility of contacts, named time-zone constraints, such as specific geographical area, time for interference, or other agency-specific reasons. The other one is to limit the contact number that a DTN node can simultaneously support, named concurrent-resources constraints, such as energy consumption, limited storage capacity, available antennas, transponders, etc. Besides these constraints, the selection criteria determined by the particular network purpose are significant to find out the most appropriate contact plan. The selection criteria which have been studied include maximum overall throughput, contact assignment fairness, traffic-driven criterion, mission-driven criterion, etc.

Algorithm Approach

The operating performance of CPD relies on a MILP theoretical model. FCP algorithm based on reduced Blossom algorithm has been designed to determine a fair contact distribution among nodes. To achieve the particular purpose of different networks, heuristics algorithms such as first improvement (FI), steepest descent (SD) and simulated annealing (SA) have been exploited to find the optimum solution from the fair contact plan delivered by FCP. FI and SD just can guarantee to find local optimum solutions; thus, to explore the complete search space, the SA was investigated. Traditional cutting-plane mechanisms were used to solve the TACP model for small space DTN networks, with obtaining an optimal solution in reasonable time. However, the

network scale expansion results in reduced computation complexity and accuracy of these methods. Therefore, approximate methodologies such as population-based meta-heuristics are worth investigating to design CPD strategy with shorter processing time and efficient output.

Key Applications

CPD is involved in efficient data forwarding to support the operations of space DTN networks. It exploits the network predictability to optimize limited resources available and improve network performance. The CPD methodologies introduced above can be used to optimize the network resources in satellite sensor networks, small satellite networks, navigation satellite networks, and Earth observation missions etc. As more information is considered in the planning stage, the rewards of CPD are more effective, while the complexity of CPD process is increasingly large.

References

- Araniti G, Bezirgiannidis N, Birrane E, Bisio I, Burleigh S, Caini C, Feldmann M, Marchese M, Segui J, Suzuki K (2015) Contact graph routing in dtn space networks: overview, enhancements and performance. *IEEE Commun Mag* 53(3):38–46
- Caini C, Cruickshank H, Farrell S, Marchese M (2011) Delay-and disruption-tolerant networking (DTN): an alternative solution for future satellite networking applications. *Proc IEEE* 99(11):1980–1997
- Fraire J, Finochietto J (2015a) Routing-aware fair contact plan design for predictable delay tolerant networks. *Ad Hoc Netw* 25:303–313. New research challenges in mobile, opportunistic and delay-tolerant networks. Energy-aware data centers: architecture, infrastructure, and communication
- Fraire JA, Finochietto JM (2015b) Design challenges in contact plans for disruption-tolerant satellite networks. *IEEE Commun Mag* 53(5):163–169
- Fraire JA, Madoery P, Finochietto JM (2013) On the design of fair contact plans in predictable delay-tolerant networks. In: *IEEE international conference on wireless for space and extreme environments*, pp 1–7
- Fraire JA, Madoery PG, Finochietto JM (2014) On the design and analysis of fair contact plans in predictable delay-tolerant networks. *IEEE Sens J* 14(11):3874–3882
- Fraire JA, Madoery PG, Finochietto JM, Leguizamn G (2015) Preliminary results of an evolutionary approach

towards contact plan design for satellite DTNs. In: 2015 IEEE international conference on wireless for space and extreme environments (WiSEE), pp 1–7

Fraire JA, Madoery PG, Finochietto JM (2016) Traffic-aware contact plan design for disruption-tolerant space sensor networks. *Ad Hoc Netw* 47:41–52

Fraire JA, Madoery PG, Finochietto JM, Leguizamn G (2017) An evolutionary approach towards contact plan design for disruption-tolerant satellite networks. *Appl Soft Comput* 52:446–456

Huang M, Chen S, Zhu Y, Wang Y (2010) Cost-efficient topology design problem in time-evolving delay-tolerant networks. In: 2010 IEEE global telecommunications conference GLOBECOM 2010, pp 1–5

Huang M, Chen S, Zhu Y, Wang Y (2013) Topology control for time-evolving and predictable delay-tolerant networks. *IEEE Trans Comput* 62(11):2308–2321

Wang Y, Sheng M, Li J, Wang X, Liu R, Zhou D (2016) Dynamic contact plan design in broadband satellite networks with varying contact capacity. *IEEE Commun Lett* 20(12):2410–2413

Zhou D, Sheng M, Wang X, Xu C, Liu R, Li J (2017) Mission aware contact plan design in resource-limited small satellite networks. *IEEE Trans Commun* 65(6):2451–2466

Contact Plan Determination

- [Contact Plan Design for Space Disruption/Delay-Tolerant Networks](#)

Content Delivery

- [Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks](#)

Content Delivery Architecture in Mobile Cellular Network

- [Mobile Content Distribution Architecture](#)

Content Distribution Architecture in Mobile Cellular Network

- [Mobile Content Distribution Architecture](#)

Content Transmission Strategy

- [Cache Placement and Content Delivery Design in Wireless Caching Networks](#)

Content-Centric Wireless Sensor Networks

- [Information-Centric Wireless Sensor Networks](#)

Contention-Based and Contention-Free Channel Access

- [Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs](#)

Control-Data Separation

- [Hyper Cellular Network: Control-Traffic Decoupled Radio Access Network Architecture](#)

Conversational Chatbot for Disaster Information

Wen-Ray Su, Ching-En Hsieh, and
Tzu-Yin Chang
National Science and Technology Center for
Disaster Reduction, Taipei, Taiwan, ROC

Synonyms

Conversational function module for disaster information with speech recognition

Definition

Applies the concepts of conversational chatbots to develop a conversational function module

that supports speech recognition in order to immediately transmit disaster warning information while also allowing users to access disaster-related information through a simple voice response system.

Historical Background

Taiwan is located in an area at high risk of natural disasters. The major natural hazards confronting Taiwan are typhoons and earthquakes. Although natural disasters cannot be avoided, losses can be effectively reduced by improving disaster resilience. As searching for disaster information becomes increasingly complex, when a disaster occurs, the community may find it difficult to immediately digest and understand complex disaster information. In order to allow people to quickly obtain and understand the information they need, easy-to-use disaster warning and information search channels ought to be provided.

With the evolution of technologies such as artificial intelligence and semantic analysis in recent years, many social media companies have developed chatbot services and have provided development modules. This concept uses text or voice conversation to operate functions. This application can analyze user semantics and understand user intentions, thereby providing the information required by the user. Automating the process of information provision can save manpower and time; users only need to be familiar with a single interface and operating procedure in order to obtain the necessary information.

Natural disasters test the resilience of government and the community. Making use of the rapid transmission of information through the Internet together with related technologies can improve disaster resilience and the communication of disaster information. Such technologies can assist us to develop easy-to-use disaster warning and disaster information channels.

On this basis, the Conversational Chatbot for Disaster Information combines the chatbot concept with disaster information, developing a con-

versational function module for disaster information with speech recognition in order to instantaneously transmit disaster warning information; this also allows users to access disaster-related information through a simple voice response system.

Since voice conversation is more natural and intuitive to operate than graphical interfaces, voice operation is increasingly valued by developers (Abdul-Kader and Woods 2015). Using voice control and computer communication, it is possible to reduce manual input time and other necessary steps; in addition, it enables users with disabilities to operate applications using voice functions (Bhargava and Nikhil 2009). For computers to understand human speech, speech recognition technology is required so the computer is able to receive and interpret the words spoken by the user.

After recognizing the natural language spoken by the user, semantic analysis is then used to understand the user's query intention. Semantic analysis needs to first establish a language understanding model, defining a combination of keywords for each possible user query intention. This model analyzes entities and classifies an intention type based on the phrases input by the user. The application can then return the information required by the user based on the results of the analysis.

The Conversational Chatbot for Disaster Information currently provides nine types of real-time data on disaster prevention and data prevention knowledge. Users can query disaster-related information in different regions according to their needs, including rainfall, weather, radar maps, landslide risk areas and warning information, water-level information, warning messages, shelters, as well as typhoon and earthquake information.

Key Applications

Providing the daily weather forecasting and disaster knowledge for public is as the general function of Conversational Chatbot for Disaster Information. When disasters are approaching,

the requests for the real-time observations, disaster alerts, and recommendations can be immediately responded with the contents of graphic information, text, and/or voice.

Microsoft's LUIS cloud semantic analysis service and Microsoft QnA Maker service are used to determine the user's query intent and a disaster knowledge quiz library, respectively. Therefore, according to various query intents, the corresponding real-time data or knowledge-based answers are delivered with messaging API to the user disaster information via graphs, texts, and/or voice.

References

- Abdul-Kader SA, Woods J (2015) Survey on chatbot design techniques in speech conversation systems. *Int J Adv Comput Sci Appl(IJACSA)* 6(7)
- Bhargava V, Nikhil M (2009) An intelligent speech recognition system for education system. <https://pdfs.semanticscholar.org/8ea7/725bab4e390e4d8ab8ebee747d2a5340c3b2.pdf>

Conversational Function Module for Disaster Information with Speech Recognition

► [Conversational Chatbot for Disaster Information](#)

Cooperative Cognitive Radio Networks

Junling Li and Weisen Shi
University of Waterloo, Waterloo, ON, Canada

Synonyms

[Cognitive radio networks](#); [Cooperative communications](#); [Cooperative sensing](#)

Definitions

Cognitive radio networks (CRNs) aim to address the issue of spectrum scarcity and meet the ever-increasing spectrum demands. In CRNs, there exist two types of users, i.e., primary users (PUs) and secondary users (SUs), where PUs with licenses can operate in certain spectrum bands, while SUs without licenses seek for access opportunities for transmissions. When cooperation is performed between users, e.g., cooperation between SUs or cooperation between SUs and PUs, the users' service quality or system performance can be significantly improved, and this type of network is referred to as cooperative cognitive radio networks.

Historical Background

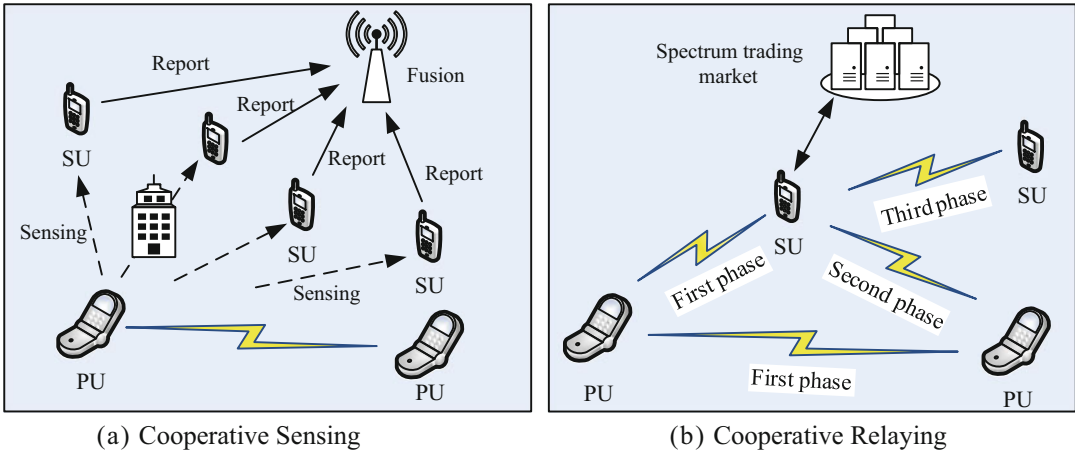
With proliferation of wireless devices and services, the demand on spectrum resources keeps increasing dramatically. As a type of natural and limited resource, spectrum is conventionally managed using a static allocation policy, where different portions of spectrum are licensed and allocated to different wireless systems for exclusive access. However, this static spectrum allocation policy can result in spectrum wastage, resulting in the issues of low spectrum utilization and spectrum scarcity. To resolve those issues and accommodate the ever-increasing spectrum demand, cognitive radio network emerges, where the SUs without licenses can opportunistically access temporally unused spectrum owned by PUs. As one of the enabling technologies, spectrum sensing proposed by Joseph Mitola in 1999 (Mitola and Maguire 1999), is to identify unused spectrum bands and protect normal operations of PUs from harmful interference. As the first standard for cognitive radio, IEEE 802.22 standard (Cordeiro et al. 2005) released in 2011 also includes spectrum sensing as a vital key component. Popular methods for spectrum sensing include matched filter detector, energy detector, cyclic feature detection, and so on (Yucek and Arslan 2009).

However, due to wireless fading and shadowing as well as limited sensing capacity, spectrum sensing performance of individual users can sometimes be unsatisfactory, adversely affecting the performance of CRNs. To address these issues, user cooperation is adopted to overcome the limitations of individual spectrum sensing, which can help better identify and explore unused spectrum for the SUs.

Foundations

Spectrum sensing is of great significance in CRNs, to detect idle spectrum bands and avoid harmful interference to PUs. Because of potential fading and shadowing, e.g., the SU might be in severe fading or shadowed by buildings, sensing errors can occur. To deal with these problems, user cooperation has been conducted in cognitive radio networks (Zhang et al. 2015). According to the participants of PUs, user cooperation can be conducted in two main forms, i.e., cooperation among SUs and cooperation between SUs and PUs. For instance, SUs can cooperatively perform spectrum sensing to improve detection performance, which is also known as cooperative spectrum sensing in the literature. Alternatively, SUs can act as cooperative relays for PUs to help enhance the PUs' transmission performance. In return, PUs grant certain access opportunities to SUs as rewards, which can be used by SUs to transmit their own data. In cases when SUs have no traffic, SUs can still cooperate with PUs to gain some rewards, such as credits, which could be spent in spectrum trading when they have data to send. Therefore, there exist various forms of cooperation in CRNs, which can be harnessed by SUs to explore spectrum access opportunities. To sum up, SUs themselves can perform cooperative spectrum sensing to better detect idle spectrum bands where PUs are idle, as shown in Fig. 1a. When PUs are detected active, SUs can act as cooperative relays for PUs to gain spectrum access opportunities or credits, depending on whether SUs have traffic to send or not, as shown in Fig. 1b.

- Cooperative spectrum sensing among SUs: Each individual SU conducts spectrum sensing and sends the results to a fusion center, where the received sensing results will be combined for making a decision using some fusion rules such as AND rule, OR rule, Majority rule, and so on (Ma et al. 2008; Zhang et al. 2014b; Akyildiz et al. 2011). Alternatively, SUs can also cooperate in a distributed way, where sensing results are exchanged among SUs to reach a consensus on spectrum availability. Cooperation among SUs can improve the detection performance for protecting PUs better and reduce the false-alarm rate for efficient use of idle spectrums, through exploiting spatial diversity and multiuser diversity.
- Cooperation with PUs for transmissions: Cooperative communication between users can improve transmission performance, such as throughput, energy efficiency, reliability, and security. Cooperation between SUs and PUs can improve the performance of PUs where SUs act as relays. As a reward, PUs grant some spectrum access opportunities for SUs' transmission in exchange for SUs' cooperation (Zhang et al. 2013; Simeone et al. 2008; Han et al. 2009; Zhang et al. 2014a). Suppose a primary system operates in a time-slotted fashion and a PU owns a period of time for transmission. The PU can divide the whole period into three portions, and SUs and PUs cooperate in a three-phase fashion accordingly. Specifically, in the first two phases, the SU acts as either a cooperative relay for enhancing the PU's performance. Then, the PU grants the last phase to the SU for access. Consequently, a "win-win" situation is achieved, where both PUs and SUs get benefits.
- Cooperation with PUs for credits: When SUs have no data to send, for the above schemes, SUs may have no incentive to collaborate with PUs. Since different parties can trade or exchange spectrum resources in spectrum trading, SUs can still gain some spectrum access opportunities in such a case. Specifically, SUs can request credits as



Cooperative Cognitive Radio Networks, Fig. 1 Cooperation in cognitive radio networks. (a) Cooperative sensing. (b) Cooperative relaying

compensations for their cooperation with PUs to improve PUs' performance. The credits earned through cooperation with PUs can be utilized by SUs later in spectrum trading (Niyato and Hossain 2008) when they have traffic.

With the above three forms of cooperation, SUs can efficiently and effectively explore spectrum access opportunities. When SUs have no data to send, through cooperation with active PUs, they can first gain a certain amount of credits and then use those credits in spectrum trading for access. When SUs have traffic to serve, SUs can collaboratively sense spectrum to identify the unused spectrum reliably and efficiently. In cases when PUs are detected active and no idle spectrum bands are available, the SUs can gain some spectrum access opportunities by serving as cooperative relays for PUs.

Key Applications

Cooperative cognitive radio networks can improve spectrum utilization and help SUs to better detect and utilize spectrum access opportunities. It can support many emerging applications or services, such as Internet of things, cognitive radio sensor networks, mobile

networks, and vehicular networks, where devices can harness cooperation and cognitive radio to accommodate large number of devices and achieve higher transmission rate.

Cross-References

- [Dynamic Spectrum Access Through Cognitive Radio Networking](#)
- [MAC in Cognitive Radio Networks](#)
- [Spectrum Sensing in Cognitive Radio Networks](#)

References

- Akyildiz IF, Lo BF, Balakrishnan R (2011) Cooperative spectrum sensing in cognitive radio networks: a survey. *Phys Commun* 4(1):40–62
- Cordeiro C, Challapali K, Birru D, Shankar S (2005) Ieee 802.22: the first worldwide wireless standard based on cognitive radios. In: *Proceedings of IEEE DySPAN 2005*
- Han Y, Pandharipande A, Ting SH (2009) Cooperative decode-and-forward relaying for secondary spectrum access. *IEEE Trans Wirel Commun* 8(10):4945–4950
- Ma J, Zhao G, Li Y (2008) Soft combination and detection for cooperative spectrum sensing in cognitive radio networks. *IEEE Trans Wirel Commun* 7(11):4502–4507
- Mitola J, Maguire GQ (1999) Cognitive radio: making software radios more personal. *IEEE Pers Commun* 6(4):13–18

- Niyato D, Hossain E (2008) Spectrum trading in cognitive radio networks: a market-equilibrium-based approach. *IEEE Wirel Commun* 15(6):71–80
- Simeone O, Stanojev I, Savazzi S, Bar-Ness Y, Spagnolini U, Pickholtz R (2008) Spectrum leasing to cooperating secondary ad hoc networks. *IEEE J Sel Areas Commun* 26(1):203–212
- Yucek T, Arslan H (2009) A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Commun Surv Tutor* 11(1):116–130
- Zhang N, Lu N, Cheng N, Mark JW, Shen X (2013) Cooperative spectrum access towards secure information transfer for CRNS. *IEEE J Sel Areas Commun* 31(11):2453–2464
- Zhang N, Cheng N, Lu N, Zhou H, Mark JW, Shen X (2014a) Risk-aware cooperative spectrum access for multi-channel cognitive radio networks. *IEEE J Sel Areas Commun* 32(3):516–527
- Zhang N, Liang H, Cheng N, Tang Y, Mark JW, Shen X (2014b) Dynamic spectrum access in multi-channel cognitive radio networks. *IEEE J Sel Areas Commun* 32(11):2053–2064
- Zhang N, Zhou H, Zheng K, Cheng N, Mark JW, Shen X (2015) Cooperative heterogeneous framework for spectrum harvesting in cognitive cellular network. *IEEE Commun Mag* 53(5):60–67

Cooperative Communications

- [Cooperative Cognitive Radio Networks](#)

Cooperative Diversity Routing

- [Opportunistic Routing \(OR\)](#)

Cooperative Localization in Wireless Sensor Networks

Xiufang Shi
Zhejiang University of Technology, Hangzhou,
China

Synonyms

[Collaborative sensor network localization](#);
[Cooperative positioning in wireless sensor networks](#); [Relative localization](#)

Definition

Cooperative localization in wireless sensor networks (WSNs) refers to estimating the sensor nodes' locations using pairwise location-related measurements. These measurements are not limited to the ones between the unknown-location nodes and the known-location nodes (i.e., anchors or beacons) and also include the ones between neighboring unknown-location nodes.

Historical Background

Cooperative localization was firstly proposed for the positioning with mobile robots (Kurazume et al. 1994) such that the positioning information can be obtained without using the landmarks in the environments. This concept was then applied in WSNs, where sensor node localization is of great significance since in many cases the sensed data become meaningful only when they are correlated with the location information. Considering that the sensor nodes are generally of low cost and low power, it is unrealistic to equip each node with the widely used Global Positioning System (GPS) module, which is too expensive and of high-energy consumption. Therefore, in many scenarios, only a small number of nodes are deployed at known locations or equipped with GPS modules, and the remaining majority nodes need to be localized. The known-location nodes, which are often called anchors, play as references for sensor node localization. Due to the short communication/sensing range of sensor nodes, most sensors have insufficient connections with anchors for location estimation. Hence, the connections between the neighboring unknown-location nodes are also utilized in localization.

Foundations

Measurements

The goal of cooperative localization is to find the locations of all the sensor nodes using the pairwise measurements from the inter-node links.

The widely used location-related measurements include (Patwari et al. 2005; Mao et al. 2007):

- *Received Signal Strength (RSS)*. In an ideal free-space environment, the received power decays in proportional to the inverse square of the transmitter-receiver distance (Rappaport et al. 1996). In real-world channels, RSS measurements are with environment-dependent errors, which mainly come from multipath signals and shadowing. RSS is typically modelled using the ensemble mean received power, which is in proportional to the logarithm of the transmitter-receiver distance. RSS measurements are of low device costs, whereas the estimation accuracy of transmitter-receiver distances from RSS measurements is low.
- *Time of Arrival (TOA)*. One-way propagation time and roundtrip propagation time measurements are both known as TOA measurements. Distance between neighboring nodes can be estimated from TOA measurement. One-way propagation time measurement requires accurate clock synchronization between the transmitter and the receiver, while roundtrip propagation time measurement has no synchronization problem. In practical environments, the accuracy of TOA measurements can be degraded by non-line-of-sight (NLOS) propagation of wireless signals.
- *Time Difference of Arrival (TDOA)*. TDOA measurement denotes the time difference of the transmitter's signal at different receivers. Distance differences between the transmitter and a number of receivers can be estimated from TDOA measurements. TDOA measurement requires clock synchronization between the receivers, and its accuracy is also affected by multipath and NLOS propagation.
- *Angle of Arrival (AOA)*. AOA measurements provide relative directional information between neighboring nodes. AOA estimation requires multiple antenna elements, which will increase the nodes' hardware cost. The accuracy of AOA measurements is affected by the directivity of the antenna, shadowing, and multipath reflections.
- *Connectivity*. The connectivity information indicates that a node is in the communication range of another node. It can define a proximity constraint between two nodes and be exploited for localization.

Localizability

Localizability means that the sensor network is uniquely localizable, i.e., all the sensor nodes in the network can be uniquely positioned in a global sense. The localizability analysis is based on graph theory by formulating the sensor network as a graph $G = (V, E)$, where V is the vertex set representing all the nodes in the network and E is the edge set representing all the inter-node links within the communication range. Each node can get location-related measurements from links to its neighbors.

For range-only network localization in two-dimensional space, if the distance measurements are noiseless, a sensor network is localizable if and only if the graph G is *generically globally rigid* and there are at least three noncollinear anchors (Aspnes et al. 2006); if the distance measurement errors are not too great, on the premise of *generically globally rigid* graph and at least three noncollinear anchors, the network will be approximated localizable, and the sensor nodes' location estimates can be near the real values (Anderson et al. 2010).

For bearing-only network localization in two-dimensional space, if the bearing measurements are noiseless, a sensor network is localizable if and only if the graph G is *parallel rigid* and there are at least two anchors; if the bearing measurement errors are small enough, on the premise of *parallel rigid* graph and at least two anchors, the network will be approximated localizable, and the sensor nodes' location estimates can be near the real values (Shames et al. 2013).

Fundamental Limits

Cramer-Rao lower bound (CRLB) is a useful tool in analyzing the fundamental limits of localization accuracy. CRLB can provide a lower bound on the variance of unbiased location estimates. It is derived from Fisher information matrix (FIM),

which can measure the amount of information that comes from the measurements. For cooperative localization in a network, FIM can be written as $\mathbf{J} = \mathbf{J}_A + \mathbf{J}_C + \mathbf{J}_P$, where $\mathbf{J}_A$, $\mathbf{J}_C$, and $\mathbf{J}_P$ respectively denote the localization information from anchors, sensors' cooperation, and a priori knowledge of the sensors' locations (Shen et al. 2010). This intuitively shows that sensors' cooperation can improve the localization accuracy. FIM for a network is a matrix of high dimensions about all the sensor nodes' locations. With regard to an individual node's location, its position error bound can be calculated using equivalent FIM (EFIM), which retains all the necessary information to determine the CRLB of this node's location (Shen and Win 2010). In cooperative localization systems, CRLB is closely related to network geometry, measurement type, measurement errors, and unknown nuisance parameters, e.g., NLOS bias. Such performance bounds can provide guidance to system designers in network deployment and measurement type selection, and can also play as benchmarks for evaluating the performance of localization algorithms.

Algorithms

Algorithms for solving the cooperative localization problem can be categorized in several different ways (Buehrer et al. 2018).

- Measurement type:* Based on the measurement being used, localization algorithms can be classified as connectivity-based, distance-based, angle-based, and hybrid methods. Connectivity-based localization (Niculescu and Nath 2001; Shang et al. 2004; Lederer et al. 2009) is of low device cost and computational complexity, while its localization accuracy is also low. Distance-based localization (Soares et al. 2015; Shi et al. 2017) and angle-based localization (Zhong et al. 2014; Lin et al. 2016) can provide more accurate location estimation, while they have higher device requirements and computational complexities. Hybrid approaches (Eren 2011; Ding et al. 2012) utilize multiple measurement types such as distance and angle, which thus
- can improve the localization accuracy of those using single measurement type.
- Centralized versus distributed:* Based on where the computation is performed, localization algorithms can be classified as centralized algorithms and distributed algorithms. In centralized algorithms (Biswas et al. 2006; Shang et al. 2004; Gholami et al. 2013), all the anchor locations and measurement data are forwarded to a processing center, and the locations of unlocalized nodes are jointly estimated at this center. In distributed algorithms (Safavi et al. 2018; Zhong et al. 2014; Zhu and Ding 2011; Soares et al. 2015; Lin et al. 2016), the location estimation is spread over the network, and each unlocalized node estimates the location by itself through local information exchange with its neighboring nodes. Centralized algorithms utilize information about the entire network and generally can provide more accurate estimates than distributed algorithms by optimizing some global objectives, e.g., maximizing the likelihood functions of the locations and minimizing the least square errors. Whereas, centralized algorithms are not robust to the processing center failure. Distributed algorithms are more scalable and robust to node failures, while they have higher requirements on the computation ability of unlocalized nodes.
- Deterministic versus probabilistic:* Based on the way of problem formulation, localization algorithms can be classified as deterministic algorithms and probabilistic algorithms. Deterministic algorithms (Biswas et al. 2006; Soares et al. 2015; Lin et al. 2016; Shi et al. 2017) assume that the locations are deterministic but unknown values; they obtain deterministic location estimates without additional information about the quality or the reliability of this solution. In probabilistic algorithms (Ihler et al. 2005; Wymeersch et al. 2009), the localization problem is formulated as a probabilistic inference problem to compute or approximate the probability distribution of each unlocalized node's location. Probabilistic algorithms can return a set of possible

location estimates preferably along with the weights quantifying the uncertainty or equivalently the reliability of the corresponding estimates. Probabilistic algorithms can provide more information about the location estimates, while their computational complexities are much higher than deterministic algorithms.

- *One shot versus tracking*: Based on whether the temporal information is utilized, localization algorithms can be classified as one shot estimation and tracking. In one shot estimation algorithms, the historical data are not utilized. In tracking algorithms (Kantas et al. 2012; Vaghefi and Buehrer 2017; Safavi et al. 2018), both the historical data and the sensors' motions are utilized. One shot algorithms can be applied to the localization of both the static sensors and mobile sensors, while tracking algorithms are generally applied in mobile sensor networks.

Key Applications

Cooperative localization plays an important role in many key applications of WSNs, such as environmental monitoring, animal tracking in biological research (Patwari et al. 2005), logistics (Patwari et al. 2005), etc. In these applications, the location information is of great significance, while GPS modules are undesirable because of high device cost and power consumption.

In recent years, with the advent of Internet of Things (IoT) and the fifth generation (5G) networks, the number of connected devices is expected to meet an exponential increase (Buehrer et al. 2018). The emerging technologies such as device-to-device (D2D) communication would enable the acquisition of D2D location-based measurements, which can be utilized in cooperative localization (del Peral-Rosado et al. 2017). Hence, IoT and 5G will result in a unique opportunity for cooperative localization (Buehrer et al. 2018). Accordingly, location-based services provided by IoT and 5G would benefit from cooperative localization.

Cross-References

► Location Based Services

References

- Anderson BD, Shames I, Mao G, Fidan B (2010) Formal theory of noisy sensor network localization. *SIAM J Discret Math* 24(2):684–698
- Aspnes J, Eren T, Goldenberg DK, Morse AS, Whiteley W, Yang YR, Anderson BD, Belhumeur PN (2006) A theory of network localization. *IEEE Trans Mob Comput* 5(12):1663–1678
- Biswas P, Lian TC, Wang TC, Ye Y (2006) Semidefinite programming based algorithms for sensor network localization. *ACM Trans Sens Netw* 2(2): 188–220
- Buehrer RM, Wymeersch H, Vaghefi RM (2018) Collaborative sensor network localization: algorithms and practical issues. *Proc IEEE* 106(6):1089–1114
- del Peral-Rosado JA, Raulefs R, López-Salcedo JA, Seco-Granados G (2017) Survey of cellular mobile radio localization methods: from 1G to 5G. *IEEE Commun Surv Tutorials* 20(2):1124–1148
- Ding G, Tan Z, Zhang L, Zhang Z, Zhang J (2012) Hybrid TOA/AOA cooperative localization in non-line-of-sight environments. In: *The 75th IEEE vehicular technology conference*. IEEE, pp 1–5
- Eren T (2011) Cooperative localization in wireless ad hoc and sensor networks using hybrid distance and bearing (angle of arrival) measurements. *EURASIP J Wirel Commun Netw* 2011(1):72
- Gholami MR, Tetrushvili L, Ström EG, Censor Y (2013) Cooperative wireless sensor network positioning via implicit convex feasibility. *IEEE Trans Signal Process* 61(23):5830–5840
- Ihler AT, Fisher JW, Moses RL, Willsky AS (2005) Nonparametric belief propagation for self-localization of sensor networks. *IEEE J Sel Areas Commun* 23(4):809–819
- Kantas N, Singh SS, Doucet A (2012) Distributed maximum likelihood for simultaneous self-localization and tracking in sensor networks. *IEEE Trans Signal Process* 60(10):5038–5047
- Kurazume R, Nagata S, Hirose S (1994) Cooperative positioning with multiple robots. In: *IEEE international conference on robotics and automation*, pp 1250–1257
- Lederer S, Wang Y, Gao J (2009) Connectivity-based localization of large-scale sensor networks with complex shape. *ACM Trans Sens Netw* 5(4):31
- Lin Z, Han T, Zheng R, Fu M (2016) Distributed localization for 2-D sensor networks with bearing-only measurements under switching topologies. *IEEE Trans Signal Process* 64(23):6345–6359

- Mao G, Fidan B, Anderson BD (2007) Wireless sensor network localization techniques. *Comput Netw* 51(10):2529–2553
- Niculescu D, Nath B (2001) Ad hoc positioning system (APS). In: IEEE global telecommunications conference, vol 5. IEEE, pp 2926–2931
- Patwari N, Ash JN, Kyperountas S, Hero AO, Moses RL, Correal NS (2005) Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal Process Mag* 22(4):54–69
- Rappaport TS et al (1996) *Wireless communications: principles and practice*, vol 2. Prentice Hall PTR, New Jersey
- Safavi S, Khan UA, Kar S, Moura JM (2018) Distributed localization: a linear theory. *Dig Proc IEEE*. <https://doi.org/10.1109/JPROC.2018.2823638>
- Shames I, Bishop AN, Anderson BD (2013) Analysis of noisy bearing-only network localization. *IEEE Trans Autom Control* 58(1):247–252
- Shang Y, Rumi W, Zhang Y, Fromherz M (2004) Localization from connectivity in sensor networks. *IEEE Trans Parallel Distrib Syst* 15(11):961–974
- Shen Y, Win MZ (2010) Fundamental limits of wideband localization part I: a general framework. *IEEE Trans Inf Theory* 56(10):4956–4980
- Shen Y, Wymeersch H, Win MZ (2010) Fundamental limits of wideband localization part II: cooperative networks. *IEEE Trans Inf Theory* 56(10):4981–5000
- Shi X, Mao G, Anderson BD, Yang Z, Chen J (2017) Robust localization using range measurements with unknown and bounded errors. *IEEE Trans Wirel Commun* 16(6):4065–4078
- Soares C, Xavier J, Gomes J (2015) Simple and fast convex relaxation method for cooperative localization in sensor networks using range measurements. *IEEE Trans Signal Process* 63(17):4532–4543
- Vaghefi R, Buehrer R (2017) Cooperative source node tracking in non-line-of-sight environments. *IEEE Trans Mob Comput* 16(5):1287–1299
- Wymeersch H, Lien J, Win MZ (2009) Cooperative localization in wireless networks. *Proc IEEE* 97(2):427–450
- Zhong J, Lin Z, Chen Z, Xu W (2014) Cooperative localization using angle-of-arrival information. In: 11th IEEE international conference on control & automation. IEEE, pp 19–24
- Zhu S, Ding Z (2011) Distributed cooperative localization of wireless sensor networks with convex hull constraint. *IEEE Trans Wirel Commun* 10(7):2150–2161

Cooperative MIMO

- Network MIMO

Cooperative Positioning in Wireless Sensor Networks

- Cooperative Localization in Wireless Sensor Networks

Cooperative Sensing

- Cooperative Cognitive Radio Networks

Cooperative Spectrum Sharing

- Matching for Cooperative Spectrum Sharing

Cooperative Transmitter-Side Interference Management

- Interference Alignment in Wireless Networks

Cooperative Wireless Networks

- Machine Learning Paradigms in Wireless Network Association

Coordinated Multipoint

- Coordinated Multipoint Transmission and Reception

Coordinated Multipoint (CoMP)

- Base Station Cooperations Under Imperfect Conditions

Coordinated Multipoint Transmission and Reception

Jian Zhao

School of Electronic Science and Engineering,
Nanjing University, Nanjing, China

Synonyms

Coordinated multipoint; CoMP

Definitions

Coordinated Multipoint (CoMP) transmission and reception comprises a series of schemes that a user equipment (UE) is served simultaneously by several points to enhance the received signal quality. A *point* refers to a set of geographically co-located transmit antennas, such as the set of antennas at a cellular base station (BS). Note that different sectors of the same BS site are thought of as different points. In CoMP, the channel state information (CSI) and/or data needs to be shared among neighboring points to coordinate their transmissions in the downlink or jointly process the received signals in the uplink. CoMP can effectively turn otherwise harmful intra-cell or inter-cell interference into useful signals and achieve significant power gain. It is an effective tool to improve the coverage of high data rates, the cell-edge throughput, and also to increase system throughput.

Historical Background

CoMP is a special technique of base station cooperation. The history of base station cooperation can be traced back to the 1990s when the concept of *macro-diversity* was proposed. In macro-diversity schemes, one or more mobile stations transfer the same message simultaneously with multiple surrounding base stations to provide diversity against long-term and short-term fading. For example, in Code Division Multiple Access

(CDMA) networks, a mobile station can communicate simultaneously with several base stations by means of soft handoff, and it can select the best channel among those base stations at any given time by selection diversity. Both coverage and capacity can be increased by means of this scheme as shown in Viterbi et al. (1994). In order to further remove the inter-cell interference penalty, researchers in academia also proposed the idea of full base station cooperation, which means all the base stations in the network are connected to a central processing site. The central processing site is responsible for collecting all the channel and data information and controls the transmission and reception of all the base stations in the network. In this way, the whole network effectively becomes a multiple-input multiple-output (MIMO) broadcast/multi-access channel with distributed antennas in the downlink/uplink, respectively. Such a network with full base station cooperation has been considered for the uplink in Hanly (1996), where the received signals at each base station are combined together before decoding at a global decoder. Thus all received signals provide useful information, and the interference at each individual base station is utilized. For the downlink, the cooperative transmission of multicell base stations was investigated in Shamai and Zaidel (2001). The results show that if the power constraints of individual base stations are ignored, the cooperative downlink transmission is recognized almost identical to that of the MIMO broadcast channel.

Research in academia has also drawn considerable interest from industry and propelled the investigation of base station cooperation in the standardization bodies, especially the Third Generation Partnership Project (3GPP). In 3GPP Long-Term Evolution (LTE) Release 8, inter-cell interference coordination (ICIC) is supported by means of coordinating message exchanges between base stations via a standardized backhaul interface named X2. The aim of ICIC is to reduce inter-cell interference for neighboring users that belong to different cells by serving those users using different frequency resources. Base stations that support this feature can generate interference information for each

frequency resource and exchange the information with neighbor base stations through the X2 interface. By learning the interference status of its neighbors, each base station can allocate the radio resources and adjust the transmit power for their users to avoid inter-cell interference. The ICIC techniques are designed primarily for semi-static coordination due to limitations of latency in the X2 interface.

The CoMP technology was introduced to standardization by 3GPP LTE-Advanced in 2011. Compared to ICIC, CoMP techniques pursue more dynamic coordination, and it is one of the core features in 3GPP Release 11 (see 3GPP TR 36.819 2013). In that specification, the CoMP scheme description, deployment scenarios, and implementation aspects, such as control signaling and UE feedback, were specified. The backhaul requirement is one of the fundamental requirements in CoMP. The discussions of CoMP in 3GPP Release 11 are mainly based on ideal backhaul assumptions. The support of CoMP involving non-ideal backhaul and the backhaul requirements for CoMP were addressed in 3GPP Release 12 (see 3GPP TR 36.874 2013). Further enhancements to CoMP operations were introduced in 3GPP Release 14 (see 3GPP TR 36.741 2017), such as support of noncoherent joint transmission.

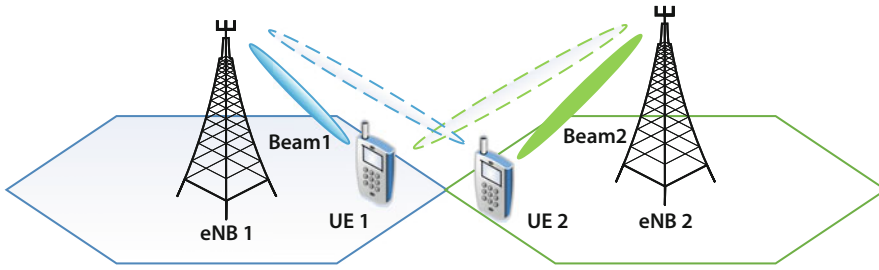
Key Applications

In traditional cellular networks, the signal quality of cell-edge users is significantly degraded compared to users in the cell center. This is mainly due to the co-channel interference from neighboring cells using the same frequency channel. CoMP transmission and reception is considered in 3GPP LTE-Advanced Release 11 as a tool to reduce inter-cell interference, boost received signal quality, and increase system throughput, especially for users at the cell edge. In 3GPP's terminology, a point is defined as a set of co-located antennas providing coverage in the same sector. Different points can refer to geographically separated base stations (BSs) or different sectors of the same BS site. The CoMP techniques dynam-

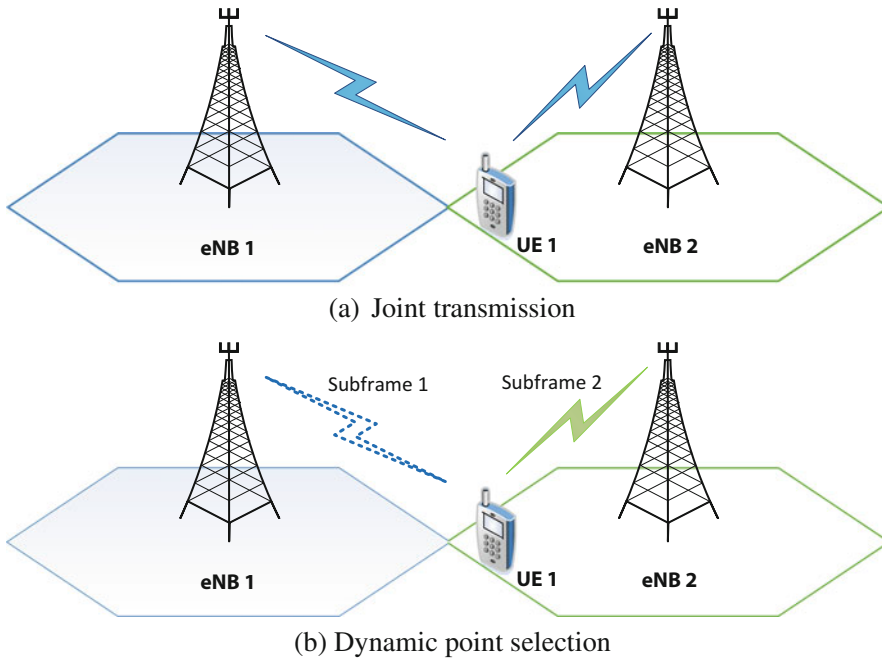
ically coordinate the transmission or reception of multiple points to provide joint scheduling of transmission as well as joint processing of received signals. In CoMP, a user equipment (UE) in the network is able to be served by two or more points. By having several BS sites or sectors to transmit or receive the same information and using specialized combining techniques, CoMP can utilize the interference constructively rather than destructively and turn the inter-cell interference (ICI) into useful signals.

In essence, the CoMP schemes can be divided into the following two major categories:

- Coordinated scheduling/coordinated beamforming (CS/CB):** In CS/CB schemes, a UE is transmitting or receiving signals to or from a single point at a given time and frequency channel. However, the communication is made with an exchange of control signals among several coordinated entities. In the downlink, CS/CB is characterized by the fact that multiple coordinated points only share channel information for the UE terminals, while the data packets that need to be conveyed to a UE terminal are available only at one point. The scheduling decision is coordinated among the transmitting points (TPs) that participate in the set of coordinating entities. When the TPs are equipped with multiple antennas, the multi-antenna TPs can adjust their beamforming coefficients to steer toward or evade certain UEs. The scheme of CB is illustrated in Fig. 1. Compared to the joint processing schemes, the CS/CB schemes have the advantage that the requirements for coordination across the backhaul network are considerably reduced. This is due to two reasons: first, each UE's data only needs to be directed to one TP; second, multiple TPs only need to coordinate scheduling decisions and details of beams. In the uplink, the scheme of coordinated scheduling minimizes interference by coordinating the scheduling decisions among the receiving points (RPs). Similar to the downlink case, different RPs coordinate with each other and only need to transfer the scheduling data.



Coordinated Multipoint Transmission and Reception, Fig. 1 Illustration of CB scheme. Solid ellipses represent intended signals. Dashed ellipses represent interferences, which are in the null space of beams



Coordinated Multipoint Transmission and Reception, Fig. 2 Joint processing schemes in the downlink. (a) Joint transmission. (b) Dynamic point selection

- **Joint processing (JP):** JP means that multiple points cooperate with each other and simultaneously transmit signals to or receive signals from UEs. For the TPs, the JP schemes can not only boost the received signal quality of the intended UEs but also actively cancel the interference to other UEs. In the downlink, there are two kinds of JP schemes for the downlink: joint transmission (JT) and dynamic point selection (DPS). In the JT scheme, two or more TPs transmit the

same message data on the same frequency and in the same subframe. When data is available for transmission at two or more TPs but only scheduled for transmission from one TP in each subframe, it is called DPS. The schemes of JT and DPS are illustrated in Fig. 2. Since multiple TPs transmit the same message signals simultaneously, the UEs' data need to be first transferred to those TPs before being transmitted to the UEs. Due to the requirement of sharing UEs' data

between multiple TPs before transmission, JP schemes place higher demand on the backhaul network (Zhao et al. 2013). In the case of uplink, it is possible to perform joint reception, which utilizes antennas at different sites to form a virtual antenna array when different BSs cooperate with each other. The signals received by different BSs are firstly exchanged through the backhaul and then combined to produce the final output signal.

In general, there is a trade-off between backhaul burden and the CoMP gain. The CS/CB schemes require less data to be transferred through the backhaul network, but resulting throughput gain is also less; the JP schemes have higher gains in network capacity at the cost of higher amount of UEs' data exchange in the backhaul.

Applications Scenarios

The applicability of the CoMP function is largely dependent on the backhaul's latency and capacity, which depends on the type of CoMP processing that can be applied and the associated performance. In order to account for the various possible network topologies and backhaul characteristics, the study of CoMP in 3GPP has focused on the following scenarios according to 3GPP TR 36.819 (2013):

- Coordination between the cells (sectors) controlled by the same macro base station, where no backhaul connection is needed;
- Coordination between cells belonging to different radio sites from a macro network;
- Coordination between a macrocell and low-power transmit/receive points within its coverage, each point controlling its own cell, i.e., with its own cell identity;
- The same deployment as the latter, except that the low-power transmit/receive points constitute distributed antennas (via RRHs) of the macrocell and are thus all associated with the macrocell identity

Cross-References

- [Base Station Cooperations under Imperfect Conditions](#)
- [Network Slicing-Enabled Green C-RAN](#)
- [Network MIMO](#)

References

- 3GPP TR 36741 (2017) Study on further enhancements to coordinated multi-point (CoMP) operation for LTE. V14.0.0
- 3GPP TR 36819 (2013) Coordinated multi-point operation for LTE physical layer aspects. V11.2.0
- 3GPP TR 36874 (2013) Coordinated multi-point operation for LTE with non-ideal backhaul. V12.0.0
- Hanly SV (1996) Capacity and power control in spread spectrum macrodiversity radio networks. *IEEE Trans Commun* 44(2):247–256
- Shamai S, Zaidel BM (2001) Enhancing the cellular downlink capacity via co-processing at the transmitting end. In: *Proceedings 53th IEEE vehicular technology conference*, Rhodes
- Viterbi AJ, Viterbi AM, Gilhousen KS, Zehavi E (1994) Soft handoff extends CDMA cell coverage and increases reverse link capacity. *IEEE J Sel Areas Commun* 12(8):1281–1288
- Zhao J, Quek TQS, Lei Z (2013) Coordinated multipoint transmission with limited backhaul data transfer. *IEEE Trans Wirel Commun* 12(6):2762–2775

Crest Factor Reduction

- [PAPR Reduction in OFDM Systems](#)

Crowd Sensing

- [Mobile Crowdsensing: Issues and Challenges](#)

Crowdsensing

- [Mobile Crowdsensing: Issues and Challenges](#)

Crowdsourced Wireless Networks

- [Sharing Economics](#)

Crowdsourcing

- [Private Truth Discovery in Cloud-Empowered Crowdsensing Systems](#)

Cyber-physical Attack

- [Cyber-physical Security in Smart Grid Communications](#)

Cyber-physical Defense

- [Cyber-physical Security in Smart Grid Communications](#)

Cyber-physical Security in Smart Grid Communications

Ruilong Deng, Peng Zhuang, and Hao Liang
Department of Electrical and Computer
Engineering, University of Alberta, Edmonton,
AB, Canada

Synonyms

[Cyber-physical attack](#); [Cyber-physical defense](#)

Definitions

Cyber-physical security in smart grid communications means to secure smart grid communications from a cyber-physical system point of view.

History Background

As widely considered to be the next generation of the power grid, smart grid has been proposed to fully upgrade the energy generation, transmission, distribution, and consumption. It can be defined as an informationized power grid that leverages information and communications technology (ICT) to automatically gather and act on meter data, in order to improve agility, reliability, efficiency, security, economy, sustainability, and environmental friendliness.

A variety of communication standards and technologies coexist in different communication networks of smart grid. Home area networks/business area networks/industrial area networks (HANs/BANs/IANs) are deployed within residential units, commercial buildings, and industrial plants for connecting multiple electrical appliances to smart meter through IEEE 802.15.4 (ZigBee), IEEE 802.11 (Wi-Fi), or power-line communication (PLC). Neighborhood area networks/field area networks (NANs/FANs) support communications among distribution substations and field electrical devices for power distribution system and microgrid operation. They connect multiple smart meters to data aggregate unit (DAU) through IEEE 802.11 (Wi-Fi), IEEE 802.16m [Worldwide Interoperability for Microwave Access (WiMax)], or cellular networks [e.g., general packet radio service (GPRS), 3G, and long-term evolution (LTE)]. Wide area networks (WANs) facilitate communications among bulk generation, power transmission system, and meter data management system (MDMS) through fiber-optic communication, microwave transmission, or cellular networks.

With the incorporation of cyber space such as ICT, smart grid becomes more open to and cyberly accessible from outside networks, such as Internet-based office networks and smart meters with two-way communication between suppliers and customers. Despite the advances, this integration also leads to new vulnerabilities of cyber-physical security, which has been reported as one of the main threats to the reliable operation of smart grid. Availability, integrity, and confiden-

tiality, also known as the CIA triad, are three of the fundamental security requirements of smart grid communications.

Confidentiality refers to the prevention of data access by unauthorized persons or entities. Smart grid communications have unintended consequences for customer privacy. Malicious attacks targeting confidentiality can be considered as privacy issues. Privacy issues have to be covered with the derived customer consumption data as they are created in metering devices. Consumption data contains detailed information that can be used to gain insights on a customer's behavior.

Integrity refers to preventing undetected modification of information by unauthorized persons or systems. For smart grid communications, this applies to information such as product recipes, sensor values, or control commands. Integrity is critical for smart grid monitoring and control, during the data collection and command notification processes, respectively. In particular, false data injection (FDI) attacks can alter the state estimation of the electrical grid and, thus, alter its operation. As a result, revenue losses are expected. On the other hand, false commands (e.g., breaker opening commands) can cause devastating effects on the electrical grid, such as large-scale blackout.

Availability refers to ensuring that unauthorized persons or systems cannot deny access or use to authorized users. For smart grid communications, this refers to all the IT elements of the plant, like control systems, safety systems, operator workstations, engineering workstations, manufacturing execution systems, as well as the communication systems between these elements and to the outside world. Malicious attacks targeting availability can be considered as denial-of-service (DoS) attacks. DoS attacks are a kind of resource consumption attacks, which send fast requests to a server or directly jam the communication channel. By compromising multiple smart meters, distributed DoS (DDoS) attacks can also be launched. Since the availability of data and commands are critical for electrical grid monitoring and control,

DoS/DDoS attack can significantly interrupt grid operation.

Key Applications

Taking FDI attacks, for example, which can circumvent bad data detection (BDD) and insert any bias into the value of the estimated state stealthily. Due to historical reasons, FDI attacks are also known as stealthy deception attacks, load redistribution attacks, malicious data attacks, data integrity attacks, and so on, proposed by different research groups at different time.

The concept of FDI attacks was first developed by Liu et al. (2009). After that, continuous works on constructing and/or defending against such attacks have been done in recent years, as a survey by Deng et al. (2017b). For the purpose of quantifying the potential threat to a power grid, two classes of security indices are introduced by Sandberg et al. (2010), corresponding to two different types of FDI attacks, namely, sparse attacks and small magnitude attacks, respectively. One of such security metrics is used by Teixeira et al. (2011) to show limitations of linear attack policies on the nonlinear AC power flow model. Besides, Teixeira et al. (2010) further propose a generalized approach to construct deception attacks on state estimation in smart grid, with specific target constraints. Dán and Sandberg (2010) consider clusters of meters at the same attack cost for the adversary to compromise and propose greedy algorithms for perfect and partial countermeasures against FDI attacks. The concept of load redistribution (LR) attacks was first introduced by Yuan et al. (2011) as a special class of FDI attacks. Kosut et al. (2011) investigate two different regimes of FDI attacks on state estimation in smart grid and investigate how FDI attacks will interfere electricity market operations, because the biased state estimation result will be used for economic dispatch. Xie et al. (2011) show that the adversary can launch FDI attacks for continuous financial arbitrage. e.g., virtual bidding at selected pairs of buses. Jia et al. (2011) consider making profit for the generator at a specific bus by launching FDI attacks on the

real-time market. Besides, Jia et al. (2012) further investigate three different scenarios: the adversary may have full, partial, or zero knowledge of real-time measurements. Bi and Zhang (2013) show that by fabricating a fake transmission congestion pattern, FDI attacks can manipulate real-time electricity price at any target bus. Bobba et al. (2010) explore how to detect FDI attacks: one way is to secure basic measurements which are selected strategically, while the other way is to verify state variables independently which are selected strategically. Kim and Poor (2011) investigate constructing FDI attacks on the power grid based on linearized measurement models and propose strategic countermeasures against such attacks, by either immunizing a small number of meter measurements or deploying phasor measurement units (PMUs). Giani et al. (2013) consider unobservable data integrity attacks on power systems and also present corresponding defense approaches by means of PMUs. Bi and Zhang (2011) propose countermeasures against FDI attacks by protecting critical state variables. In addition, Bi and Zhang (2014) further propose a mixed protection strategy, in case that either fails to obtain the defense objective. Göl and Abur (2013) identify the vulnerability of state estimation against cyberattacks and provide two PMU-based countermeasures, by either converting critical measurements to redundant ones or eliminating the leveraging effect of leverage measurements. Deng et al. (2017a) assume that whether or not a meter measurement could be compromised by an adversary depends on the defense budget deployed by the defender on the meter and design a least-budget defense strategy from this perspective, which is suitable for large-scale power systems.

As a typical cyber-physical system, smart grid communications are vulnerable to a wide variety of cyber and/or physical attacks. In literature, both cyber and physical attacks have been investigated. In addition to aforementioned cyberattacks, (Salmeron et al. 2004; Holmgren et al. 2007; Chen et al. 2011) have discussed the optimal allocation of defense resources against physical attacks. Recently, to demonstrate the vulnerability of smart grid to joint cyber and

physical attacks, Soltan et al. (2015) made the first attempt to consider an adversary disconnecting some transmission lines and blocking related information to the control center. Li et al. (2016, 2018) showed that cyberattacks could mask transmission line outages, even for local attacks with incomplete network information. Deng et al. (2017c) investigate coordinated cyber-physical attacks (CCPAs) in smart grid and present two new kinds of CCPAs, for which cyberattacks can accurately mask transmission line outages by replaying meter readings and utilizing PMU measurements, respectively. Countermeasures are developed against the two kinds of CCPAs, respectively, based on known-secure PMU measurement verification and the observation that cyberattacks cannot mask the impact of physical attacks on the physical space (i.e., the power system equivalent impedance) which can be online tracked. This research indicates that the vulnerability and security of smart grid communications must be revisited from a cyber-physical system point of view.

Cross-References

- [Fault Tolerance and Reliability of Smart Grids](#)
- [Privacy-Preserving Data Aggregation for Smart Grids](#)

References

- Bi S, Zhang YJ (2011) Defending mechanisms against false-data injection attacks in the power system state estimation. In: Proceedings of IEEE GLOBECOM Workshops (GC Wkshps), pp 1162–1167
- Bi S, Zhang YJ (2013) False-data injection attack to control real-time price in electricity market. In: Proceedings of IEEE global communications conference (GLOBECOM), pp 772–777
- Bi S, Zhang YJ (2014) Using covert topological information for defense against malicious attacks on DC state estimation. *IEEE J Sel Areas Commun* 32(7):1471–1485
- Bobba RB, Rogers KM, Wang Q, Khurana H, Nahrstedt K, Overbye TJ (2010) Detecting false data injection attacks on DC state estimation. In: Preprints of the First Workshop on Secure Control Systems, CPSWEEK

- Chen G, Dong ZY, Hill DJ, Xue YS (2011) Exploring reliable strategies for defending power systems against targeted attacks. *IEEE Trans Power Syst* 26(3):1000–1009
- Dán G, Sandberg H (2010) Stealth attacks and protection schemes for state estimators in power systems. In: *Proceedings of IEEE international conference on smart grid communications (SmartGridComm)*, pp 214–219
- Deng R, Xiao G, Lu R (2017a) Defending against false data injection attacks on power system state estimation. *IEEE Trans Ind Inf* 13(1):198–207
- Deng R, Xiao G, Lu R, Liang H, Vasilakos AV (2017b) False data injection on state estimation in power systems—attacks, impacts, and defense: a survey. *IEEE Trans Ind Inf* 13(2):411–423
- Deng R, Zhuang P, Liang H (2017c) CCPA: coordinated cyber-physical attacks and countermeasures in smart grid. *IEEE Trans Smart Grid* 8(5):2420–2430
- Giani A, Bitar E, Garcia M, McQueen M, Khargonekar P, Poolla K (2013) Smart grid data integrity attacks. *IEEE Trans Smart Grid* 4(3):1244–1253
- Göl M, Abur A (2013) Identifying vulnerabilities of state estimators against cyber-attacks. In: *Proceedings of the IEEE PowerTech*, pp 1–4
- Holmgren ÅJ, Jenelius E, Westin J (2007) Evaluating strategies for defending electric power networks against antagonistic attacks. *IEEE Trans Power Syst* 22(1):76–84
- Jia L, Thomas RJ, Tong L (2011) Malicious data attack on real-time electricity market. In: *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp 5952–5955
- Jia L, Thomas RJ, Tong L (2012) Impacts of malicious data on real-time price of electricity market operations. In: *Proceedings of the IEEE Hawaii international conference on system sciences (HICSS)*, pp 1907–1914
- Kim TT, Poor HV (2011) Strategic protection against data injection attacks on power grids. *IEEE Trans Smart Grid* 2(2):326–333
- Kosut O, Jia L, Thomas RJ, Tong L (2011) Malicious data attacks on the smart grid. *IEEE Trans Smart Grid* 2(4):645–658
- Li Z, Shahidehpour M, Alabdulwahab A, Abusorrah A (2016) Bilevel model for analyzing coordinated cyber-physical attacks on power systems. *IEEE Trans Smart Grid* 7(5):2260–2272
- Li Z, Shahidehpour M, Alabdulwahab A, Abusorrah A (2018) Analyzing locally coordinated cyber-physical attacks for undetectable line outages. *IEEE Trans Smart Grid* 9(1):35–47
- Liu Y, Ning P, Reiter MK (2009) False data injection attacks against state estimation in electric power grids. In: *Proceedings of the ACM conference on computer and communications security (CCS)*, pp 21–32
- Salmeron J, Wood K, Baldick R (2004) Analysis of electric grid security under terrorist threat. *IEEE Trans Power Syst* 19(2):905–912
- Sandberg H, Teixeira A, Johansson K (2010) On security indices for state estimators in power networks. In: *Preprints of the First Workshop on Secure Control Systems, CPSWEEK*
- Soltan S, Yannakakis M, Zussman G (2015) Joint cyber and physical attacks on power grids: graph theoretical approaches for information recovery. In: *Proceedings of the ACM SIGMETRICS*, pp 1–14
- Teixeira A, Amin S, Sandberg H, Johansson KH, Sastry SS (2010) Cyber security analysis of state estimators in electric power systems. In: *Proceedings of IEEE Conference on Decision and Control (CDC)*, pp 5991–5998
- Teixeira A, Dán G, Sandberg H, Johansson KH (2011) A cyber security study of a SCADA energy management system: Stealthy deception attacks on the state estimator. In: *IFAC World Congress*, vol 18, pp 11271–11277
- Xie L, Mo Y, Sinopoli B (2011) Integrity data attacks in power market operations. *IEEE Trans Smart Grid* 2(4):659–666
- Yuan Y, Li Z, Ren K (2011) Modeling load redistribution attacks in power systems. *IEEE Trans Smart Grid* 2(2):382–390

Cyber-physical System Security

► IoT Security

Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation

Xiqing Liu
Department of Engineering Science, National Cheng Kung University (NCKU), Tainan City, Taiwan

Synonyms

Insufficient CP-OFDM/OFDMA systems; OFDM/OFDMA system without CP; Reduced guard interval OFDM/OFDMA

Definitions

CP-free OFDM/OFDMA system refers to an orthogonal frequency-division multiplexing

(OFDM) system or an orthogonal frequency-division multiple access (OFDMA) system, without insertion of cyclic prefix (CP) at the transmitter.

History Background

With an increasing demand for a higher data rate, such as 5G, more bandwidths will be required, which makes transmit signals extremely sensitive to inter-symbol interference (ISI) in a multipath channel. Orthogonal frequency-division multiplexing (OFDM) or orthogonal frequency-division multiple access (OFDMA), working as a multi-carrier transmission scheme, receives much attention due to its ability to combat ISI. A common approach to tackle ISI for an OFDM/OFDMA system is to insert a cyclic prefix (CP) at the beginning of every block before transmission. As long as the length of CP is made longer than channel delay spread, ISI will only occur within the CP period. After CP removal at the receiver, the data can be recovered successfully. However, the CP-based OFDM/OFDMA has to pay an expensive cost, that is, a loss of spectral efficiency, which will be one of the most serious limitations in futuristic high-rate wireless applications. In 3GPP LTE standard (3GPP TS 36.104 V13.2.0 2016), about 25% of an OFDM/OFDMA block is consumed by extended CP, and this is definitely a heavy price for the increasingly scarce spectrum resource.

To increase the spectral efficiency, the simplest method is to prolong the length of an OFDM/OFDMA block. The block length is determined by the symbol (such as the QAM symbol and the PSK symbol, etc.) time and the number of subcarriers. Prolonging the symbol time will reduce data rate, whereas increasing the number of subcarriers may result in a serious peak to average power ratio (PAPR) problem (Goldsmith 2005) and a high complexity in FFT/IFFT implementation (Sohn 2014), and what is more important is that the length of an OFDM/OFDMA block is limited by the channel's coherent time. Generally, the minimum

interval between two pilot signals is the block length. If the block length is longer than the coherence time, it will be difficult to obtain the accurate channel state information (CSI) in time via the existing channel estimation techniques. As a result, the receiver will suffer from a serious error when performing the coherent detection. In terms of analysis given above, it can be known that CP insertion will surely become a performance bottleneck in future OFDM/OFDMA systems due to its significant loss in spectral efficiency, and the CP-free OFDM/OFDMA addressed in this section is believed to be more competitive than CP-OFDM/OFDMA in future wireless applications.

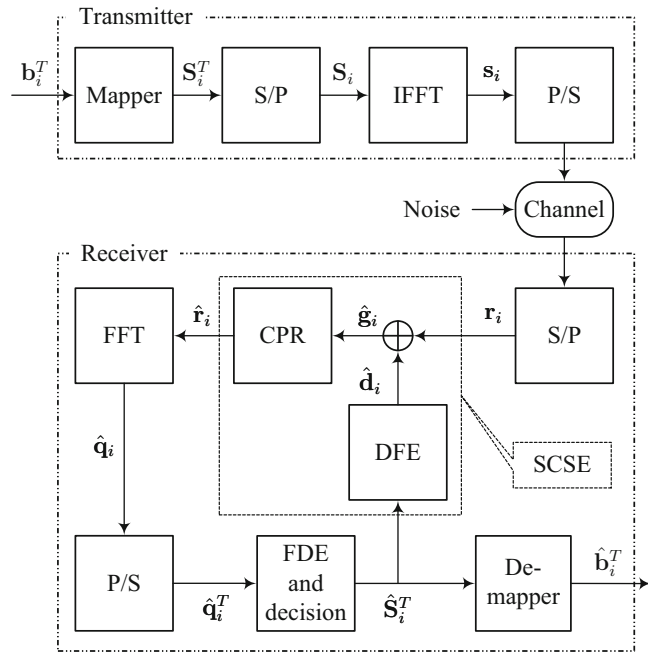
The spectral efficiency degradation caused by CP insertion poses a challenge to the traditional OFDM/OFDMA systems and attracts lots of efforts to solve it. As the length of a CP is mainly decided by the length of channel impulse response (CIR), many works investigated CIR shortening schemes as an effort to reduce the CP length. More details can be found in Zhang et al. (2003), Zhang and Qin (2010), Meng et al. (2008), and Lindqvist and Fertner (2011). Also, CIR may be different in various links and may change with time, and hence adjusting CP length to the current channel conditions is beneficial in terms of achievable data rate. In Tonello et al. (2010) and Huang and Rao (2012), an online CP adaptation method was developed for determining optimal CP length via estimating the channel statistics during transmission, while in this section, two kinds of CP-free OFDM algorithms will be introduced in detail, i.e., symbolic cyclic shift equalization (SCSE) and time domain precoding (TDP).

Symbol Cyclic Shift Equalization

Figure 1 is the block diagram of the SCSE-OFDM. It can be seen that a quadrature amplitude modulation (QAM) mapper maps the i th binary input $\mathbf{b}_i^T$ into a block of QAM symbols $\mathbf{S}_i^T$. Then, $\mathbf{S}_i^T$ is sent into a serial to parallel (S/P) converter before IFFT module. Thereafter, an IFFT is performed on $\mathbf{S}_i$ to generate an OFDM

**Cyclic Prefix-Free
OFDM/OFDMA Systems,
Design,
and Implementation,**

Fig. 1 This is the block diagram of SCSE-OFDM system, including its transmitter and receiver



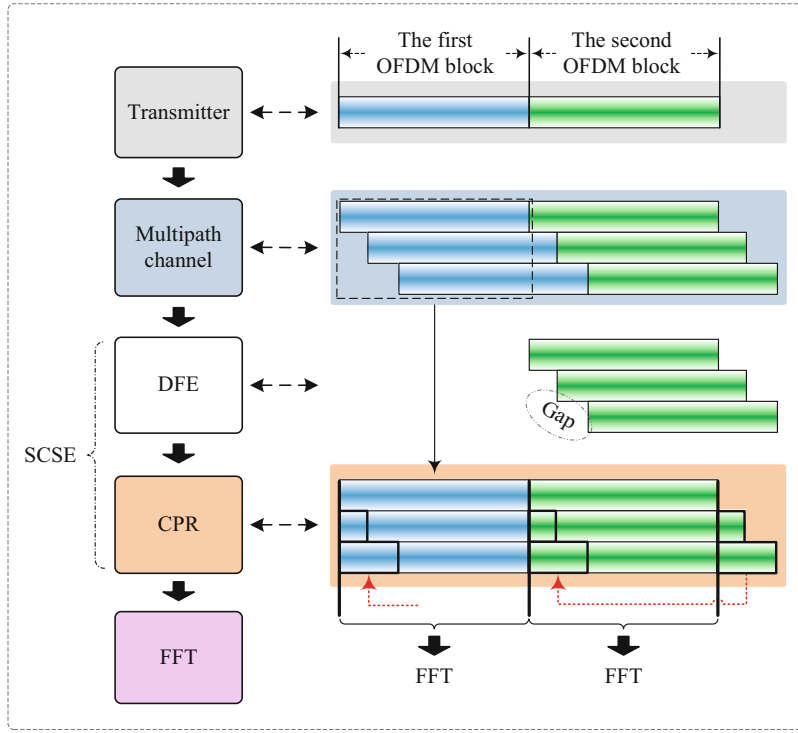
block, i.e., s_i , and then s_i go through a parallel to serial converter (P/S). After that, s_i^T will be sent into the channel. Here the baseband discrete multipath channel is considered.

Having experienced multipath propagation, the i th block of the received signal is r_i^T . Send r_i^T to S/P to get r_i , and then perform SCSE algorithm on r_i before FFT. SCSE is formed by decision feedback equalization (DFE) and CP regeneration (CPR). Specifically, DFE is implemented first to suppress ISI, yielding $\hat{g}_i$, and then $\hat{g}_i$ is fed into CPR to get $\hat{r}_i$. After SCSE, the newly constructed block will be sent to FFT and P/S to get $\hat{q}_i^T$. To remove the flat fading gain, a one-tap frequency domain equalization (FDE) needs to be implemented on $\hat{q}_i^T$ before decision device. After making the decision, a QAM demapper translates the recovered QAM symbols $\hat{S}_i^T$ into recovered binary bits $\hat{b}_i^T$.

Through the description given above, SCSE algorithm is the most concern, and Fig. 2 presents the main idea. Assume two consecutive blocks are considered, and the second block is the interest. It can be seen after multipath propagation, the ISI is the tail of the previous block, and it will be eliminated by the DFE, leaving a “Gap.” CPR pads this “Gap” before FFT. Here the block

is called the first OFDM block as it is preceded by a null period, whose length is at least equal to channel’s delay spread, and thus the interference from its previous blocks cannot be seen by the current OFDM block. However, after DFE, the different linear shifts induced by multipath propagation produce a “Gap” in the multipath return, which will destroy the orthogonality between the subcarriers if feed the block to FFT directly (Malkin et al. 2008). The reason is that FFT/IFFT is a cyclical algorithm for time/frequency conversion, which will lead to unrecovered distortion if the input blocks are linearly shifted in its processing interval (Lyons 2011). Thus, although the ISI is removed from the block, CP still should be generated. The process to generate the CP is called CP regeneration (CPR), and DFE plus CPR is named as symbol cyclic shift equalization (SCSE). With the help of SCSE, the linearly shifted blocks can be turned into cyclically shifted blocks.

Without loss of generality, assume the i th ($i = 2, 3, \dots$) OFDM block is the interest. As shown in Fig. 1, SCSE starts with the DFE procedure for ISI elimination. Assume when performing SCSE algorithm on the i th OFDM block, the detection of the $(i - 1)$ th OFDM block has



Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation, Fig. 2 A schematic diagram with respect to SCSE algorithm with two consecutive OFDM blocks

been completed. Therefore, with the CSI known by the receiver, the tail of the $(i - 1)$ th OFDM block can be reconstructed, which is defined as $\hat{\mathbf{d}}_i$ in Fig. 1. Mathematically, the DFE can be expressed by,

$$\hat{\mathbf{g}}_i = \mathbf{r}_i - \hat{\mathbf{d}}_i. \quad (1)$$

Theoretically, ISI can be eliminated completely with the help of DFE, if no CSI error is introduced. However, after DFE, a “Gap” is produced between the neighboring OFDM blocks. As mentioned earlier, the DFE-processed “Gap” will result in serious ICI after the blocks have been sent into FFT, and padding the “Gap,” termed as CPR, is an important step.

For simplicity, the signal processing of the CPR algorithm will be illustrated with the help of an example presented in Fig. 3. Assume an OFDM block contains four symbols, which expe-

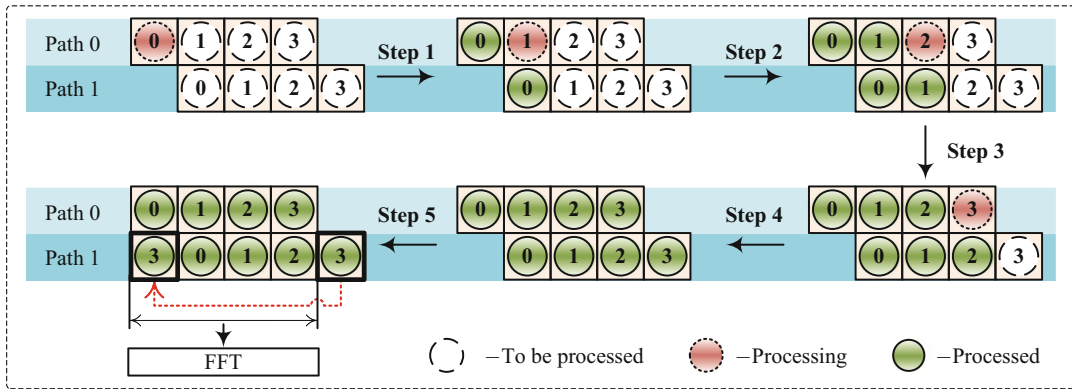
riences a two-ray multipath channel. The receiver is assumed to achieve synchronization with the first path to arrive, and the second block is the multipath return, whose delay equals to one sample time.

Step 1: The channel gains are assumed to be known by the receiver, and thus the first sample is utilized to estimate the first symbol in Path 1.

Step 2: Subtract “①” from the second sample, and the “①” in Path 0 can be got. Divide it by channel gain of Path 0, and then multiply by the channel gain of Path 1 to achieve the estimation of “①” in Path 1. In this way, all the values of “①” and “②” can be acquired.

Step 3: Having acquired “①,” subtract it from third sample to yield the “②” in Path 0, which is then used to estimate “②” in Path 1.

Step 4: In this step, “②” will be subtracted from the fourth sample, and the estimation of “③” in Path 0 will be got. Next, the estimated “③” is



Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation, Fig. 3 The signal processing of an example with respect to CPR algorithm. Here “ $\textcircled{n}$ ” ($n = 0, \dots, 3$) refers to the n th symbol in an OFDM block

used to estimate the symbol beyond this detection interval, i.e., the fourth symbol in Path 1.

Step 5: After Steps 1–4, the estimations in all of the balls can be available. Specifically, the fourth symbol in Path 1, i.e., “ $\textcircled{3}$,” exceeds the detection interval, which can be regarded as the CP signal and will be fed into the “Gap” to construct the cyclically shifted OFDM block. Finally, send the new-constructed OFDM block into the follow-up FFT processor.

It can be observed from Fig. 3 that all the balls will be taken into the calculation to estimate CP signal. If an OFDM block contains a larger number of symbols, more calculation steps will be required. Therefore, to perform CPR, its high computational complexity will be a challenge for practical applications. To tackle this issue, Liu et al. (2017b) develop a low complexity algorithm for CPR implementation by constructing the conjugation symmetrical OFDM block at the transmitter.

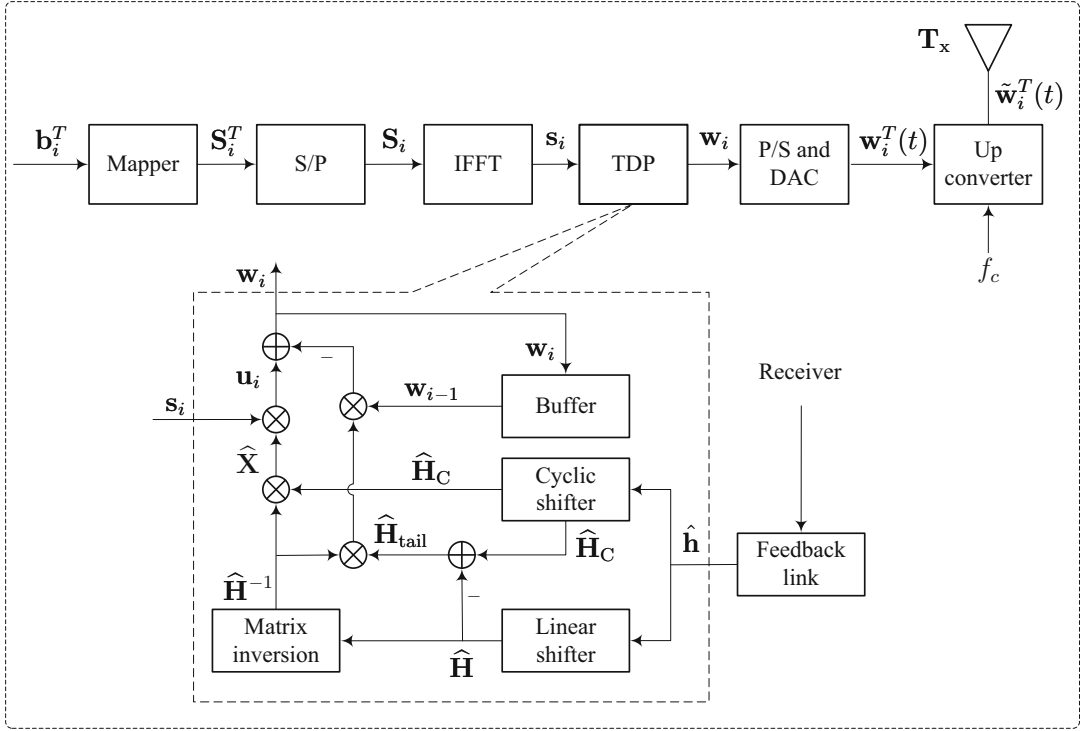
Though SCSE can help an OFDM system to remove CP in a multipath channel, it has some drawbacks. For the DFE scheme, the decision error will inevitably occur due to fading, noises, impaired CSI, etc. If the decision error is conveyed to the “Gap,” the polluted gap will spread the error to the detection of following blocks, degrading the bit error rate (BER) performance. In addition, in some scenarios, where the power

of the first path to arrive is relatively weak, the estimation error produced by CPR can also degrade the BER performance since the gain of the first path to arrive always serves as the denominator to estimate the symbols in other paths. Moreover, detection delay caused by SCSE poses a challenge for the hardware, especially in a high data rate scenario. To deal with these issues, in the next subsection, a time domain precoding (TDP) algorithm designed for CP-free OFDM system will be introduced, and the mathematical derivations will be also given to validate its feasibility.

Time Domain Precoding

Figure 4 is the block diagram of the TDP-OFDM. As shown in Fig. 4, $\mathbf{b}_i^T$ is the binary bit input. QAM mapper will map the binary bit sequence into a block of QAM symbols, defined as $\mathbf{S}_i^T$. Next, $\mathbf{S}_i^T$ is fed into S/P and IFFT to generate $\mathbf{s}_i = \mathbf{F}^{-1}\mathbf{S}_i$, in which $\mathbf{F}^{-1}$ denotes the IFFT matrix (Lyons 2011). After IFFT, perform TDP algorithm on $\mathbf{s}_i$ to get $\mathbf{w}_i$. The precoding matrix $\hat{\mathbf{X}}$ is formed by,

$$\hat{\mathbf{X}} = \hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_{\mathbf{C}} \quad (2)$$



Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation, Fig. 4 This is the transmitter block diagram of the proposed TDP-OFDM system

where the matrices of $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}_C$ can be acquired by linearly shifting and cyclically shifting the vector of $\hat{\mathbf{h}}$. Specifically, $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}_C$ can be expressed by,

$$\hat{\mathbf{H}} = \begin{bmatrix} \hat{h}(0) & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \hat{h}(L-1) & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \hat{h}(L-1) & \cdots & \hat{h}(0) \end{bmatrix}_{M \times M} \quad (3)$$

and

$$\hat{\mathbf{H}}_C = \begin{bmatrix} \hat{h}(0) & 0 & \cdots & 0 & \hat{h}(L-1) & \cdots & \hat{h}(1) \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \hat{h}(L-2) & \ddots & \ddots & & \ddots & \ddots & \hat{h}(L-1) \\ \hat{h}(L-1) & \ddots & \ddots & & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \hat{h}(L-1) & \hat{h}(L-2) & \cdots & \hat{h}(0) \end{bmatrix}_{M \times M} \quad (4)$$

where $\hat{\mathbf{h}} = [\hat{h}(0), \cdots, \hat{h}(L-1), 0 \cdots, 0]^T$ is the CSI vector with length extended to M and $\hat{h}(l)$ channel gain of Path l ($l = 0, \cdots, L-1$). Here M and L ($L < M$) refers to the lengths of the block and the CIR, respectively. It can be seen that $\hat{\mathbf{H}}_C$ is a cyclic matrix with its columns shifted circularly (Meyer 2000). Due to the property of a cyclic matrix, we can have

$$\mathbf{F}\hat{\mathbf{H}}_C\mathbf{F}^{-1} = \hat{\Lambda} = \begin{bmatrix} \hat{\lambda}_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\lambda}_{M-1} \end{bmatrix} \quad (5)$$

in which $\mathbf{F}$ and $\mathbf{F}^{-1}$ denote FFT and IFFT matrix. $[\hat{\lambda}_0, \dots, \hat{\lambda}_{M-1}]$ represents the frequency response of $\mathbf{h}$ (Chen et al. 2009). To perform precoding scheme on $\mathbf{s}_i$, we will first make a multiplication to get $\mathbf{u}_i$, or $\mathbf{u}_i = \hat{\mathbf{X}}\mathbf{s}_i$. In Fig. 4, $\hat{\mathbf{H}}_{\text{tail}}$ refers to the tail of the channel matrix induced by multipath propagation, which is calculated via

$$\hat{\mathbf{H}}_{\text{tail}} = \hat{\mathbf{H}}_C - \hat{\mathbf{H}} = \begin{bmatrix} 0 & \cdots & 0 & \hat{h}(L-1) & \cdots & \hat{h}(1) \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & & \ddots & \hat{h}(L-1) \\ \vdots & & & \ddots & & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}_{M \times M} \quad (6)$$

Next, use $\hat{\mathbf{H}}_{\text{tail}}$ and the buffer's output to *pre-eliminate* the interference from the transmit signal, and we can have

$$\mathbf{w}_i = \mathbf{u}_i - \hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_{\text{tail}}\mathbf{w}_{i-1}, \quad (7)$$

which can be divided into two parts, i.e., information signal ($\mathbf{u}_i$) and the interference elimination signal ($\hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_{\text{tail}}\mathbf{w}_{i-1}$). As shown in Fig. 4, after S/P and digital to analog converter (DAC), $\mathbf{w}_i$ will be modulated by radio-frequency carrier and then transmitted into channel through antenna $\mathbf{T}_x$.

The transmit signal experiences the multipath channel to arrive at the receiver, and we can have the i th received OFDM block as

$$\mathbf{r}_i = \mathbf{H}\mathbf{w}_i + \Upsilon_{i-1} + \mathbf{v}_i, \quad (8)$$

where $\mathbf{v}_i$ is the noise term and Υ_{i-1} is the ISI term. In a multipath channel, the interference is caused by the tail of the previous block, i.e.,

$$\Upsilon_{i-1} = \mathbf{H}_{\text{tail}}\mathbf{w}_{i-1}, \quad (9)$$

in which $\mathbf{w}_{i-1}$ denotes the $(i-1)$ th OFDM block, and with the result in (Eq. 7), $\mathbf{H}\mathbf{w}_i$ can be further written as

$$\begin{aligned} \mathbf{H}\mathbf{w}_i &= \mathbf{H}[\mathbf{u}_i - \hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_{\text{tail}}\mathbf{q}_{i-1}] \\ &= \mathbf{H}\mathbf{u}_i - \mathbf{H}\hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_{\text{tail}}\mathbf{w}_{i-1}. \end{aligned} \quad (10)$$

Substituting Eqs. 9 and 10 into Eq. 8 yields

$$\begin{aligned} \mathbf{r}_i &= \mathbf{H}\mathbf{u}_i - \mathbf{H}\hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_{\text{tail}}\mathbf{w}_{i-1} + \mathbf{H}_{\text{tail}}\mathbf{w}_{i-1} + \mathbf{v}_i \\ &\triangleq \mathbf{H}\mathbf{u}_i + \mathbf{v}_i. \end{aligned} \quad (11)$$

It is noted that the second equality holds up under the assumption the perfect CSI can be fed back to the transmitter. Then, send $\mathbf{r}_i$ to the FFT processor, and we can get

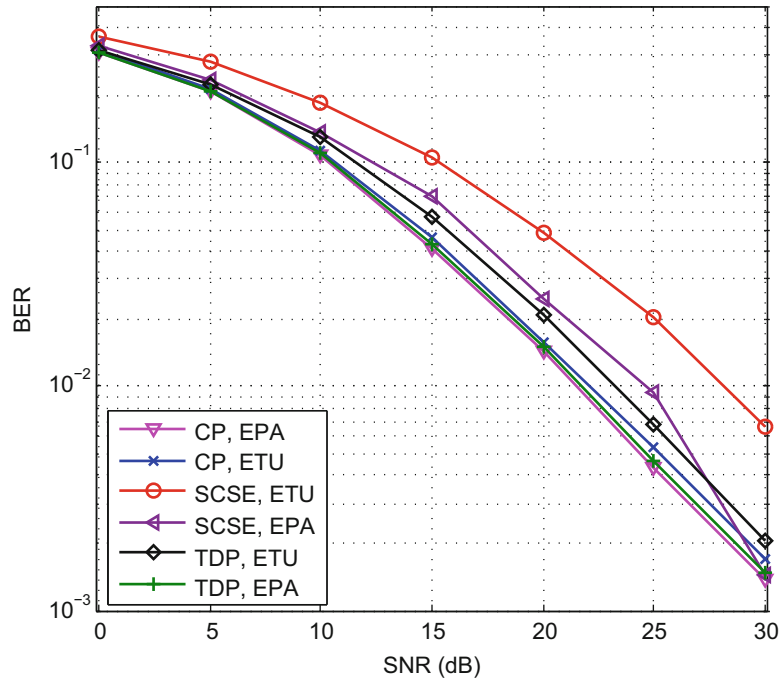
$$\mathbf{q}_i = \mathbf{F}[\mathbf{H}\mathbf{u}_i + \mathbf{v}_i] = \Lambda\mathbf{S}_i + \mathbf{F}\mathbf{v}_i \quad (12)$$

where $\mathbf{F}$ denotes the FFT matrix. After P/S, $\mathbf{q}_i^T$ is sent into one-tap FDE, decision device, and demapper in turn. Finally, we can get the recovered binary data $\hat{\mathbf{b}}_i^T$.

Performance Evaluation

In Fig. 5, the 3GPP LTE channel models were employed to evaluate the BER performance (3GPP TS 36.104 V13.2.0 2016), where the modulation type is 16QAM and the number of subcarriers is 1024. The channel models are extended typical urban (ETU) and extended pedestrian A (EPA). With the bandwidth of 100 MHz, the delay spreads of ETU and EPA channels are 500 and 41, respectively. For the CP-OFDM system, the CP length is made equal to delay spread. From Fig. 5 we can observe that TDP-OFDM and CP-OFDM have a similar BER performance, but BER performance of SCSE-OFDM is significantly poor, especially under the ETU scenario, where the first path is a relatively weak (3GPP TS 36.104 V13.2.0 2016) and CPR may bring considerable estimation error. On the other hand, CP-OFDM overcomes multipath

Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation,
Fig. 5 The BER is the transmitter block diagram of the proposed TDP-OFDM system



interference by means of CP insertion, but paying a price of spectral efficiency degradation. In the comparison of spectral efficiency, SCSE-OFDM and TDP-OFDM are better than CP-OFDM. With 16QAM modulation, the spectral efficiency of TDP-OFDM or SCSE-OFDM is about 4 bps/Hz in either ETU channel or EPA channel. However, the spectral efficiencies of CP-OFDM in the ETU and EPA channels are about 2 bps/Hz and 3.8 bps/Hz, respectively.

The number of complex multiplication is used here to evaluate the computational complexity of SCSE and TDP. For SCSE, $\frac{M}{2}\log_2 M + ML$ complex multiplications are consumed in the DFE (Liu et al. 2017a; Cooley and Tukey 1965). Also, the number of complex multiplications needed by CPR is ML . Thus, to complete the detection of the i th OFDM block, the total number of complex multiplications required by SCSE is

$$\mathcal{O}_{\text{SCSE}} = M \left(\frac{\log_2 M}{2} + 2L \right). \quad (13)$$

Different from SCSE, the computational complexity of the TDP algorithm is mainly allocated at the transmitter. The number of complex mul-

tiplications required for inversion transform of $\hat{\mathbf{H}}$ is $M \log_2 M$ (Commenges and Monsion 1984). As shown in Eq. 2, $\hat{\mathbf{H}}^{-1}$ is multiplied by $\hat{\mathbf{H}}_C$ to construct the precoding matrix, producing the computational complexity of M^3 . Besides, the calculation of $\hat{\mathbf{X}} \times \mathbf{s}_i$ will require M^2 complex multiplications, and the calculation presented in Eq. 7 will cost $M^3 + M^2$ complex multiplications. Therefore, to detect the i th OFDM block, the total number of complex multiplications consumed by the TDP algorithm is

$$\mathcal{O}_{\text{TDP}} = 2(M^3 + M^2) + M \log_2 M. \quad (14)$$

Be aware that the computational complexity of TDP is related to the channel coherence time. The calculations of $\hat{\mathbf{H}}^{-1}\hat{\mathbf{H}}_C$ and $\hat{\mathbf{H}}_{\text{tail}}\hat{\mathbf{X}}$ are just needed to be performed once within the same coherent time. Generally, the length of an OFDM block is much shorter than coherence time, and thus Eq. 14 holds an upper bound for TDP's complexity.

This contribution introduces two CP-free OFDM/OFDMA algorithms, i.e., SCSE and TDP. It is clear that the spectral efficiency of

the CP-free OFDM/OFDMA outperforms that of a traditional CP-OFDM/OFDMA. TDP is better than SCSE with respect to the BER performance owing to the decision error and the estimation error caused by DFE and CPR, respectively. However, TDP requires CSI must be fed back to the receiver timely to construct the precoding matrix, which poses a serious challenge to design a powerful feedback link, especially in a rapidly time-varying scenario. Moreover, TDP is power inefficient if it works in a deep fading channel (Goldsmith 2005). Therefore, a definite answer to which scheme is the optimal option is not available, as each scheme has its unique strength for a particular case.

Key Applications

At present, OFDMA has been widely studied and was chosen as the standard in the 3GPP LTE downlink multiple access scenario (3GPP TS 36.104 V13.2.0 2016). However, the CP degrades the spectral efficiency significantly, especially in a long delay spread channel; CP-free OFDM/OFDMA technology will be more competent than traditional CP-OFDM/OFDMA under the case where the spectral is the most concern. Additionally, CP-free OFDM/OFDMA systems may achieve more considerations in latency-sensitive wireless applications. Latency reductions in OFDM/OFDMA systems require substantial reductions in the block length, while CP holds the critical limitation to shorten an OFDM/OFDMA block, whose length is decided by the delay spread. Fortunately, in the CP-free OFDM/OFDMA systems introduced in this contribution, the multipath interference can be solved by equalization or precoding instead of CP, and the length of an OFDM/OFDMA block can be shortened without the limitation caused by CP. For this occasion, the CP-free OFDM/OFDMA may be a more proper choice than CP-OFDM/OFDMA in a latency-sensitive applications.

Cross-References

- [Analog and Digital Communications](#)
- [Multiple Access Technique for Cellular Wireless Networks](#)

References

- 3GPP TS 36104 V1320 (2016) Evolved universal terrestrial radio Access (E-UTRA); base station (BS) radio transmission and reception (release 13). <https://www.3gpp.org/>
- Chen Z, Chang Y, Liu F, Zhou J, Yang D (2009) A turbo FDE technique for OFDM system without cyclic prefix. In: 2009 IEEE 70th vehicular technology conference fall, pp 1–5
- Commenges D, Monsion M (1984) Fast inversion of triangular toeplitz matrices. *IEEE Trans Autom Control* 29(3):250–251
- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex fourier series. *Math Comput* 19(90):297–301
- Goldsmith A (2005) *Wireless communications*. Cambridge University Press, New York, pp 381–382
- Huang Y, Rao BD (2012) Awareness of channel statistics for slow cyclic prefix adaptation in an OFDMA system. *IEEE Wireless Commun Lett* 1(4):332–335
- Lindqvist F, Fertner A (2011) Frequency domain echo canceller for DMT-based systems. *IEEE Signal Process Lett* 18(12):713–716
- Liu X, Chen HH, Chen S, Meng W (2017a) Symbol cyclic-shift equalization algorithm – A CP- free OFDM/OFDMA system design. *IEEE Trans Veh Technol* 66(1):282–294
- Liu XQ, Chen HH, Lv BY, Meng WX (2017b) Symbol cyclic shift equalization PAM-OFDM – a low complexity CP-free OFDM scheme. *IEEE Trans Veh Technol* 66(7):5933–5946
- Lyons RG (2011) *Understanding digital signal processing*. Prentice Hall, Boston
- Malkin M, Hwang CS, Cioffi JM (2008) Transmitter precoding for insufficient-cyclic-prefix distortion in multicarrier systems. In: IEEE vehicular technology conference, pp 1142–1146
- Meng L, Li J, Huang A, Song J, Zhang H (2008) MMSE channel shortening equalization for OFDM systems with unequal subcarrier powers. In: Third international conference on communications and networking in China, pp 1013–1017
- Meyer C (2000) *Matrix analysis and applied linear algebra*, society for industrial and applied mathematics, pp xxii,714
- Sohn I (2014) A low complexity PAPR reduction scheme for OFDM systems via neural networks. *IEEE Commun Lett* 18(2):225–228

- Tonello AM, D'Alessandro S, Lampe L (2010) Cyclic prefix design and allocation in bit-loaded OFDM over power line communication channels. *IEEE Trans Commun* 58(11):3265–3276
- Zhang P, Qin J (2010) A simple channel shortening equalizer for wireless TDD-OFDM systems. In: 2010 IEEE international conference on wireless communications, networking and information security, pp 83–86
- Zhang J, Ser W, Zhu J (2003) Effective optimisation method for channel shortening in OFDM systems. *IEE Proc Commun* 150(2):85–90

D

D2D Communication in Cellular Systems

- [Resource Management in Macrocell–Small Cell Systems and D2D-Assisted Cellular Systems](#)

Data Analytics for Longitudinal Biomedical Data

Hua Fang
University of Massachusetts (UMass)
Dartmouth, North Dartmouth, MA, USA
UMass Medical School, Worcester, MA, USA

D2D Communications

- [Neighborhood-Based ICT-Driven Support \(NIS\) System of Emergency Medical Services \(EMS\) for Elderly People in Hyper-Aged Societies](#)

Synonyms

[Data streams](#); [Longitudinal data](#); [Time series](#)

Definition

Longitudinal data, a common data type in the biomedical area, refer to the data where each participant has repeatedly measured values on the same variable at two or more time points. Time series data have more time points for each participant, (e.g., hourly, daily, weekly and monthly, etc.) compared to regular longitudinal data. Data streams refer to the real-time (on-the-fly) millisecond data for each participant generated from wireless biomedical devices such as sensors. Broadly speaking, time series and data streams are all longitudinal data with the difference in the frequency of time points when collecting data for each participant.

Data Aggregation

- [Application of Machine Learning in Wireless Sensor Network](#)

Data Analytics

- [Data-Driven Pattern Prediction](#)

Background for Longitudinal Biomedical Data Analytics

Common longitudinal data analytical models have been well studied in the statistical and biostatistical fields, ranging from common linear and nonlinear regression models (e.g., mixed or hierarchical models, mixture models) (McCulloch et al. 2008; Muthén et al. 2002; Skrondal and Rabe-Hesketh 2004) to emerging statistical learning (or machine learning) methods (e.g., lasso, random forests, artificial neural networks, MIFuzzy) (Murphy 2012; James et al. 2013; Hastie et al. 2009; Fang 2017).

Biomedical data can be small or big in terms of sample size. Big data are termed differently, called large-scale or high-dimensional data in statistics. Many emerging methods have been developed to analyze and understand these big data in the areas of genome, biomedical images, behavioral and administrative health data studies, and web-based and wireless health (Bishop 2006; Rajpurkar et al. 2017; Fang et al. 2015).

Key Applications of Data Analytics for Longitudinal Biomedical Data

Regularized approaches such as lasso (L1) and ridge (L2) are increasingly applied in longitudinal biomedical data, when the number of variables is large and potential overfitting is likely. Bias and variance trade-off issue needs to be considered, and usually the model reducing a substantial amount of variance is recommended even if a small amount of bias is introduced during the modeling process (e.g., Murphy 2012).

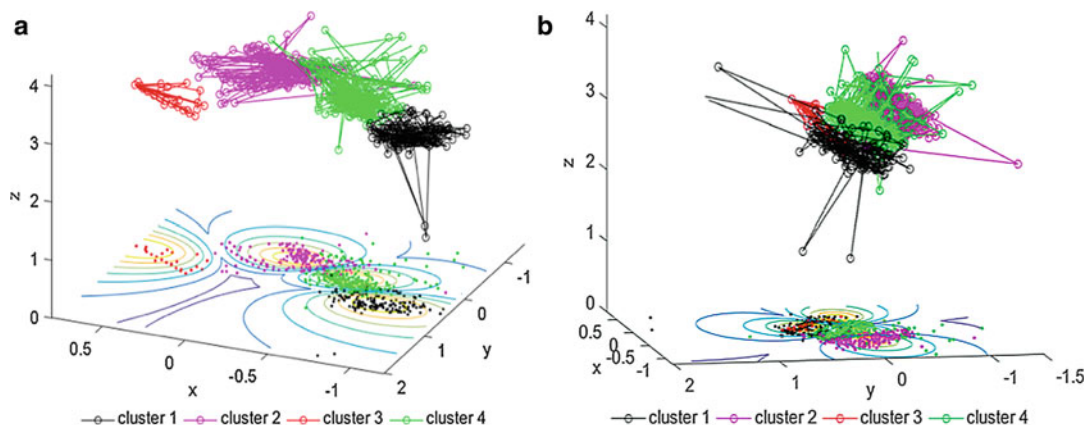
Validation is imperatively needed for biomedical data analytics. Numeric and graphic indices have been devised for specific methods, such as mixed and mixture models. Novel statistical learning methods are being developed for more complex longitudinal biomedical data, but not all have included vigorous validation procedure. Cross-validation, bagging, bootstrapping, and boosting methods are available and should be

applied in biomedical data analytics (Efron and Tibshirani 1993; James et al. 2013; Hastie et al. 2009; Zhang et al. 2016b).

Visualization methods such as those used in trajectory pattern recognition methods cannot just better present the high-dimensional longitudinal data but also assist in analyzing and validating patterns in biomedical data. More visual analytics in longitudinal studies are being developed. A visualization-aided validation has been studied to help behavioral trajectory pattern recognition of big longitudinal data (e.g., Fang and Zhang 2017; Zhang et al. 2016a). Figure 1 demonstrates that, by optimizing view angles, optimized 3D projection can better visualize patterns or clusters, compared to a random 3D projection (Fig. 1a, b). Figure 1a shows best 2D projection from optimized 3D view angle, while Fig. 1b indicates a poor 2D projection from random 3D view angle.

Multiple imputation-based fuzzy learning (MIFuzzy) is one of the most recently developed soft computing methods, used for learning the trajectory patterns of small or big complex data (e.g., high-dimensional, pattern overlapping, and error prone) with missing values in longitudinal random controlled trials and observational studies, including e-Health and mobile health (Fang 2017; Fang et al. 2009, 2010, 2011, 2012; Zhang and Fang 2016; Zhang et al. 2016a, b; Kim et al. 2017). MIFuzzy provides a comprehensive validation procedure, including fuzzy-logic based numeric indices, visual analytics, and simulation tools. The graphical user interface software for MIFuzzy is available online. It can be integrated with regularized approaches and other supervised learning methods such as neural networks. MIFuzzy is now being developed into a semi-supervised fuzzy learning method for prediction in biomedical studies (e.g., Gurugubelli et al. 2019).

Computational methods for real-time biomedical data streams are being developed, but these data are mostly downloaded and analyzed offline, as “historical” real-time data. Well-known ECG and EKG data streams are well studied, but data analytics for irregular behavioral data streams



Data Analytics for Longitudinal Biomedical Data, Fig. 1 (a) Best 2D projection from optimized 3D view angle. (b) Poor 2D projection from random 3D view angle

and multiple data streams are underdeveloped. Parameter trajectory description method based on descriptive statistics and slide window combined with supervised statistical learning methods such as K-nearest means and random forests are proposed for irregular behavioral data streams and multiple concurrent data streams which measure different bio-physiological parameters (Carreiro et al. 2015, 2016). However, a gap still exists between the conceptual promise of these approaches and the development of computationally efficient algorithms for the real on-the-fly event detection in these biosensor systems.

Conclusion

Common statistical models have been widely applied in biomedical studies. As more complex biomedical data are available, new statistical learning methods are being developed. Regularized approaches, validation, and visualization are key components for biomedical data analytics. Substantial efforts are needed to develop fast and accurate algorithms for analyzing behavioral and multiple concurrent biomedical data streams.

References

- Bishop C (2006) Pattern recognition and machine learning. Springer, New York. (Online) ISBN: 13: 9780387310732
- Carreiro S et al (2015) iMStrong: deployment of a biosensor system to detect cocaine use. *J Med Syst* 39(12):186. <https://doi.org/10.1007/s10916-015-0337-9>. PMID:26490144
- Carreiro S et al (2016) Wearable biosensors to detect physiologic change during opioid use. *J Med Toxicol* 12(3):255–262. PMID:27334894
- Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York
- Fang H (2017) MIFuzzy clustering for incomplete longitudinal data in smart health. *Smart Health (Elsevier Journal)* 1–2:50–65. <https://doi.org/10.1016/j.smhl.2017.04.002>. PMID:28993813. PMCID:PMC5631-546. <https://www.umassmed.edu/fanglab/research-projects/>. Accessed 25 July 2019
- Fang H, Zhang Z (2017) An enhanced visualization method to aid behavioral trajectory pattern recognition infrastructure for big longitudinal data. *IEEE Trans Big Data*. Accepted. <https://doi.org/10.1109/TBDATA.2017.2653815>. NIHMSID:907513
- Fang H et al (2009) Pattern recognition of longitudinal trial data with nonignorable missingness: an empirical case study. *Int J Inf Technol Decis Mak* 8(3):491–513. PMID:20336179
- Fang H et al (2010) A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm. *Pattern Recogn* 43(4):1393–1401. PMID:20300543

- Fang H et al (2011) A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. *Neurotoxicol Teratol* 33(1):155–165. PMID: 21256430
- Fang H et al (2012) Detecting graded exposure effects: a report on an East Boston pregnancy cohort. *Nicotine Tob Res* Sep 14(9):1115–1120. PMID: 22266824
- Fang H et al (2015) A survey on big data research. *IEEE Netw Mag* 29(5):6–9. <https://doi.org/10.1109/MNET.2015.7293298>. PMID: 26504265
- Gurugubelli VS et al (2019) Neuro-Fuzzy classifier for longitudinal behavioral intervention data. 2019 international conference on computing, networking and communications (ICNC): cloud computing and big data. <https://doi.org/10.1109/ICCNC.2019.8685574>
- Hastie T et al (2009) *The elements of statistical learning*. Springer. (Online) ISBN: 978-0-387-84858-7
- James G et al (2013) *An introduction to statistical learning*. ISBN: 978-1-4614-7138-7 (online) and 978-1-4614-7137-0
- Kim SS et al (2017) Acculturation, depression, and smoking cessation: a trajectory pattern recognition approach. *Tob Induc Dis*. <https://doi.org/10.1186/s12971-017-0135-x>. PMCID: PMC5525352
- McCulloch CE et al (2008) *Generalized, linear, and mixed models*, 2nd edn. Wiley, Hoboken
- Murphy KP (2012) *Machine learning: a probabilistic perspective*. The MIT Press, Cambridge, MA. ISBN: 978-0-262-01802-9
- Muthén B et al (2002) General growth mixture modeling for randomized preventive interventions. *Biostatistics* 3(4):459–475
- Rajpurkar P et al (2017) CheXNet: radiologist-level pneumonia detection on chest X-Rays with deep learning. <https://arxiv.org/pdf/1711.05225.pdf>. Accessed 25 July 2019
- Skrondal A, Rabe-Hesketh S (2004) *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, New York
- Zhang Z, Fang H (2016) Multiple- vs non- or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data. In: *Proceedings of 2016 IEEE first conference on connected health: applications, systems and engineering technologies* (NIH & NSF jointly sponsored). <https://doi.org/10.1109/CHASE.2016.19>
- Zhang Z et al (2016a) A new MI-based visualization aided validation index for mining big longitudinal web trial data. *IEEE Access* 4:2272–2280. <https://doi.org/10.1109/ACCESS.2016.2569074>. NIHMSID:790905. PMID: 27482473. PMCID: PMC4963037
- Zhang Z et al (2016b) Multiple imputation based clustering validation (MIV) for big longitudinal trial data with missing values in eHealth. *J Med Syst* 40(6):1–9. <https://doi.org/10.1007/s10916-016-0499-0>. PMID: 27126063. PMCID: PMC4881752

Data Center Network in Cloud Computing

Jiaju Qi and Long Zhao
BUPT, Beijing, China

Synonyms

Cloud computing; DCN

Definitions

Data center is composed of multiple parts, including a set of complex computer server equipment, the system of data communication connection, the system of data storage, the equipment of environment control and energy providing, the monitoring and safety equipment, etc. It is a collection of not only a series of computer server equipment but also a complex system running large-scale service applications, which can deal with large-scale computing data for business and organization.

DCN is the network applied to data center, which is composed of data centers and the connections among them. These connections include the structure of network topology, the equipment of routing and switching, and the protocols of connections. As to data center, its inner traffic can present typical features such as the increasing and high concentration of exchange data flow. Thus, the following requirements are needed in DCN, including ability on large-scale data, high expansibility, high robust, low configuration overhead, high bandwidth and efficient network protocol between the servers, flexible topology structure, link capacity control, traffic isolation, green energy saving and low cost, etc.

Cloud computing is an Internet-based computing mode. Based on cloud computing, information and resources of hardwares and softwares are available to computer servers and other devices (Wang et al. 2008). A typical cloud provider

provides a common network business application service, which can be interviewed by browser software.

Historical Background

If we only consider traditional data center, which is composed of computing module, storage module, communication module, and so on, DCN has a long history of development and can date from ENIAC. Then with the rise of cloud computing, SaaS (Kavis 2014) has realized the transition from the need for computing resources brought about by the infrastructure to the on-demand ordering model. DCN with cloud computing function has provided users with massive growth in data bandwidth resources. There are some DCN architectures designed for cloud computing. The tree-based DCNs like fat-tree and three-tier have been used broadly at last in the enterprise cloud (Fox et al. 2009). However they lack scalability (Meng et al. 2010), based on which some new DCN architectures have been proposed, such as fat-tree (Al-Fares et al. 2008) and VL2 (Greenberg et al. 2009).

Foundations

- The use of DCN in cloud computing:
 - Advantages

Since softwares and data are stored on servers in the data center, computing power can be used as a commodity to circulate via the Internet. As to cloud computing, DCN is an indispensable infrastructure of cloud computing data center. There are three advantages of DCN cloud computing services:

 1. DCN can effectively connect a large number of data centers to implement cloud computing services by applying some ingenious topology structure.
 2. DCN provides high communications between large-scale computer server equipment.
 3. DCN supports virtualization infrastructure, such as virtual machines, virtual switches, and virtual links, which gives DCN high scalability. The isolation and migration of the virtual infrastructure are very convenient for DCN.
 - Demands

In recent years, the demand and scale of cloud computing have increased rapidly. However, there are some shortcomings of DCN established by traditional networks, including the following problems: Infranet cannot carry mass traffic, Infranet is not fast enough, as well as the scalability of the network is inadequate. Application of DCN to cloud computing services should satisfy both availability and scalability. There are two new trends in the development of the DCN: the connection of DCN needs to be more simplified, and the service provided by DCN needs to be more stable.
 - Structures of DCNs for cloud computing

DCN for cloud computing uses a three-tier separate architecture from top to bottom, namely, cloud computing layer, virtual layer, and physical layer. Thereinto, we focused on cloud computing layer:

 - Cloud computing layer

The function of cloud computing layer provides an important foundation for cloud activities. In this layer, important modules such as big data, cloud service model, and cloud application can be implemented and applied. At the same time, the cloud computing layer also needs to monitor the virtual layer below in real time and allocate resources for the virtual layer.
 - Cloud service model

The cloud service model is a directly user-oriented module, which has three common definitions: infrastructure as a service, namely, IaaS; platform as a service, namely, PaaS; and software as a service, namely, SaaS (Kavis 2014):

1. Services to customers by IaaS are the applications run by operators on cloud computing infrastructure. Users can access it on a variety of devices through a client interface, such as a browser. Consumers do not need to manage or control any cloud computing infrastructure, including networks, servers, operating systems, storage, and so on.
2. Services to consumers by PaaS are the applications developed by customers using the development languages and tools provided.
3. Services to consumers by SaaS are utilized for all computing infrastructures, including CPU, internal storage, networks, and other basic computing resources, and users are able to deploy and run arbitrary software, including operating systems and applications.

For cloud service model, DCN has two challenges. One is that DCN needs centralized deployment and dynamic adjustment of network resource allocation, which leads to low efficiency. Second, it is difficult to charge in this business model.

– Big data analysis

Big data analysis and processing of cloud applications is also an important module of cloud computing layer. At present, DCN needs to support rapid processing of large amounts of data based on users' reliance on cloud demand. There are three requirements for DCN, the size of data, namely, volume; the size of data types, namely, velocity; and the speed of data processing, namely, variety. Volume requires that the nodes of DCN have abundant computing resources, including hardware resources, which can hold large amounts of data storage. Velocity requires that the nodes of DCN have data portability to complete tasks in different working environments. And variety requires that DCN's communication links have larger bandwidth to support high-

throughput data communications (Wang et al. 2015).

• Topology of DCNs for cloud computing

The design of topology structure for cloud computing needs to be improved for relieving central link blockage, expanding the network, and increasing end-to-end communication efficiency. There are three physical DCN architectures worth watching: hierarchical model, recursive model, and rack-to-rack model (Hammadi and Mhamdi 2014):

– Hierarchical model

Each layer of the hierarchical model uses different methods to control traffic flow by switches, which can effectively reduce network congestion. Among them, the most commonly used is three-tier DCN, namely, core, aggregation, and edge layers. The data flow starts from edge layer, passes through aggregation layer, and flows to core layers. Thanks to three-tier DCN's stable structure and ease of implementation, its use is more common. In addition, there are also fat-tree (Al-Fares et al. 2008), VL2 (Greenberg et al. 2009), and some other hierarchical architectures. They reduce the traffic load overhead at core layer to some extent and have high cost-efficiency. Hierarchical model is currently the most used DCN architecture, but it has some problems. One is that when the network is too large, the topology of hierarchical model architecture is too redundant, resulting in low scalability of the whole network.

– Recursive model

A switch and many servers constitute a cell, and the recursive model is composed of different cells. Different cells are connected by servers that perform better than normal ones and have higher computing speeds. Compared with hierarchical model, recursive model fully takes into account the importance of the implementation of high scalability, that is, the data center needs to be extended. In addition, recursive model has a higher utilization ratio than hierarchical model on the switch, resulting in lower cost and higher efficiency.

The structure of recursive model facilitates the implementation of load balancing. However, the performance of the recursive model depends too much on the bridge server, which leads to hard implementation in practice.

- Wang L, Tao J, Kunze M, Castellanos AC, Kramer D, Karl W (2008) Scientific cloud computing: early definition and experience. In: 10th IEEE international conference on high performance computing and communications, 2008, HPCC'08. IEEE, pp 825–830
- Wang B, Qi Z, Ma R, Guan H, Vasilakos AV (2015) A survey on data center networking for cloud computing. *Comput Netw* 91:528–547

Key Applications

In the future, the development of IoT cloud computing will be more and more significant. As a necessary part of the cloud computing infrastructure, DCN for cloud computing has become a bridge among data centers and undertake the function of network computing and network storage.

Cross-References

- [Data Center Network Virtualization](#)
- [Edge Data Centers in Fog Computing](#)
- [Load Balancing in Data Center Networks](#)

References

- Al-Fares M, Loukissas A, Vahdat A (2008) A scalable, commodity data center network architecture. In: ACM SIGCOMM computer communication review, vol 38. ACM, pp 63–74
- Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I (2009) Above the clouds: a Berkeley view of cloud computing. In: Department of electrical engineering and computer sciences. University of California, Berkeley, Report UCB/EECS 28(13):2009
- Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S (2009) V12: a scalable and flexible data center network. In: ACM SIGCOMM computer communication review, vol 39. ACM, pp 51–62
- Hammadi A, Mhamdi L (2014) A survey on architectures and energy efficiency in data center networks. *Comput Commun* 40:1–21
- Kavis MJ (2014) Architecting the cloud: design decisions for cloud computing service models (SaaS, PaaS, and IaaS). Wiley
- Meng X, Pappas V, Zhang L (2010) Improving the scalability of data center networks with traffic-aware virtual machine placement. In: INFOCOM, 2010 proceedings IEEE. IEEE, pp 1–9

Data Center Network Virtualization

Xiong Xiong¹ and Rongtao Xu²

¹Beijing University of Posts and Telecommunications, Beijing, China

²State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Synonyms

- [Network virtualization in the data center](#)

Definitions

A data center (DC) is a facility used to house computer systems and associated components, such as servers (physical machines), storage and network devices (e.g., switches, routers, and cables), power distribution systems, and cooling systems.

A data center network (DCN) is the communication network that interconnects computer systems and associated components in a data center. Generally, a data center network can be described by the network topology, routing/switching equipments, and the network protocols.

Network virtualization is the process of combining hardware and software network resources and network functionality into a single, software-based administrative entity, a virtual network.

The data center network virtualization is a term that extends virtualization concepts such as abstraction, pooling, and automation to the data center network. Virtualization usually refers to

hardware virtualization, which typically applies the software or firmware called hypervisor to create multiple isolated and independent virtual machines (VMs) upon the underlying hardware resources (i.e., physical machine). Similar to hardware virtualization, network virtualization aims at creating multiple virtual networks (VNs) on top of a shared physical network substrate allowing each VN to be implemented and managed independently. A virtual network consists of a set of virtual networking resources, such as virtual nodes (including end hosts, switches, routers) and virtual links.

Historical Background

The conventional DCN is organized in a three-tier architecture that follows a multi-rooted tree-based network topology, where network switches are divided into three layers, namely, access, aggregation, and core layers, from the bottom up (Bari et al. 2013). As the lowest layer, the access layer consists of top-of-rack (ToR) switches that provide direct connectivity to the servers mounted on every rack. The switch in the aggregation layer is called aggregation switch (AS). Each AS forwards traffic from multiple access layer ToR switches to the core layer, while core layer switches are responsible for connecting the data center to the Internet. To improve the reliability of DCN, the adjacent layers are commonly connected with multiple links for redundancy.

A classic three-tier topology used in the DCN is fat-tree topology, which is a special type of Clos topology (Bari et al. 2013). The fat-tree topology is built of k power of deliveries (PODs), and each pod contains $k/2$ ASs and $k/2$ ToR switches. Each ToR switch is directly connected to $k/2$ servers, i.e., each POD contains $(k/2)^2$ servers. The core layer contains $(k/2)^2$ core switches where each of the core switches is connected to one aggregate layer switch in each of the PODs. The fat-tree topology can apply identical and cheap commodity products for all switches, and provide full bisection bandwidth in theory. However, the scalability is one of the major issues due to the limited number of ports

in each switch. Except for fat-tree topology, there are many other tree-based topologies adopted in the DCN, such as VL2, Aspen tree, Jellyfish, etc. (Xia et al. 2017). Moreover, the server-centric topology also plays an important role in DCN, e.g., DCell, BCube, and Scafida (Xia et al. 2017).

The conventional DCN is often constructed using layer-2 and layer-3 network, in which layer-2 networking is used within a POD (i.e., between access and aggregation layers) and layer-3 networking is used between PODs (i.e., between aggregation and core layers). In layer-2 networking, all switches in a POD form a layer-2 broadcast domain, which may lead to loop due to multipath redundancy. To deal with the problem, various spanning tree protocols (xSTP) are used to eliminate loops and prevent layer-2 broadcast storms, allowing only one of the redundant links to be used to forward traffic (Xia et al. 2017). Although this method reduces bandwidth efficiency, this layer-2 and layer-3 network architecture is the mainstream and works well in the conventional data center.

However, with the advent of cloud computing era, virtualization technology (commonly achieved by server virtualization) is widely used in data center. It is because virtualization technology can improve server utilization, allocate resources on demand, and reduce operation and maintenance costs. A physical machine can be virtualized to multiple VMs, and each VM runs independently with different MAC and IP addresses. Thus, data centers are required to support hundreds of thousands of servers (both physical and virtual machines). The main limitation of traditional DCN is scalability since commodity switches were not designed to handle a large number of VMs. In particular, switches have to maintain an entry in their forwarding information bases (FIBs) for every VM, which can dramatically increase the size of forwarding tables. Moreover, DCNs are also required to deliver high cross-section bandwidth to support a high volume of east-to-west traffic for cloud computing applications. Therefore, the traditional layer-2 and layer-3 network architecture fails to meet DCN requirements in the cloud computing era due to its low bandwidth efficiency.

In virtualized data center, a new requirement is that dynamic VM migration must be supported to maximize service reliability and deployment flexibility. However, VMs can only be migrated within a POD on layer-2 network. It is because if VMs are migrated across a layer-2 domain, their IP addresses have to be reassigned, which will make applications interrupted. Therefore, the need for a layer-2 network extension is required to support potentially millions of servers (VMs) in DCN, where VM migration can carry out in a large domain without being limited by the size of POD. Moreover, the non-blocking and low-latency data forwarding is required, and fat-tree topology must be supported to make optimal use of multiple data forwarding paths between switches. In this situation, the traditional layer-2 networking using xSTP protocols for a loop-free topology can lead to a large number of disabled links and result in performance deterioration.

Supporting multi-tenancy is another requirement for data center. A physical data center is commonly shared by multiple tenants. Each tenant has its own virtual data center consisting of a groups of VMs, storage and network resources. For the sake of security, traffic between tenants needs to be isolated. In traditional layer-2 networking, isolation is usually achieved through VLANs by dividing the network into multiple layer-2 broadcast domains. However, the number of supported tenants is limited by the number of VLANs defined in IEEE 802.1Q, i.e., a maximum of 4096 (IEEE 2014).

As can be seen above, the traditional DCN using layer-2 and layer-3 network architecture with xSTP and VLAN protocols cannot meet the scalability and management needs of large-scale virtual data center. Therefore, data center network virtualization emerges to deal with the problems.

Foundations

The foundations of data center network virtualization are various network virtualization techniques and network management techniques (Jain and Paul 2013; Alshaer 2015; Tatsuhiko et al.

2013). Network virtualization techniques can be divided into two classes, i.e., device virtualization and link virtualization (Wang et al. 2015).

Device virtualization refers to abstract physical hardware into multiple logical (virtual) switches and routers. A virtual switch (vSwitch) is a software application that acts as a layer-2 hardware switch providing communication between VMs and physical machines. vSwitches are usually implemented in hypervisors that run on physical machines to provide virtual environment for VMs. The key role of vSwitch is responsible for forwarding data packets between virtual network interface cards (NICs) of VMs and the physical NICs of host machine. That is, multiple NICs share the physical NICs with the help of vSwitch. The most common method to implement vSwitch is the virtual Ethernet bridge (VEB) and virtual Ethernet port aggregator (VEPA) techniques (IEEE 2012b). Both VEB and VEPA are designed to handle local VM-to-VM traffic, but they focus on different purposes. A VEB is a VLAN bridge aiming at switching the traffic of multiple VMs over an internal virtual network, while a VEPA forces the traffic of one or more VMs to be handled by an external switch. Moreover, VEPA is a simple extension to VEB that address many of the limitations with VEBs, such as traffic invisible and not configurable. An enhancement to VEPA is multichannel VEPA, in which Q-in-Q tagging mechanism is applied (IEEE 2006).

Another kind of device virtualization is to merge multiple physical switches at the same network layer on the logical switch without changing the existing physical network topology. This kind of technique brings many advantages, such as simplifying network structure, facilitating network protocol deployment, and improving network reliability and manageability. Benefit from it, the number of logical links between the logical switches can be reduced. Thus, the logical loops can be prevented, although loops may exist in the underlying physical network. However, this kind of technique is implemented privately by different vendors. Therefore, the products of different vendors cannot interoperate with each other. Some typical technologies are Huawei's cluster switch system (CSS), intelligent

stacking (iStack) and super virtual fabric (SVF), Cisco's virtual switching system (VSS) and fabric extender (FEX), H3C's intelligent resilient framework (IRF), VMware's vSphere distributed switch (VDS), etc.

Just as its name implies, link virtualization refers to the technologies that establish logical links between virtual network devices, e.g., VMs, vSwitches, etc. The key point of link virtualization is to use encapsulation techniques (e.g., tagging or tunneling) to build up overlay networks on the underlying physical network. The key objectives are to achieve location-independent addressing, scaling up the number of logical networks, providing network isolation for multi-tenant environment, etc. There are various levels of link virtualization technologies, which can be partitioned into several classes, such as layer 2 over layer 2 (L2oL2), layer 2 over layer 3 (L2oL3), etc.

There are various L2oL2 techniques, including provider bridge (PB, also known as Q-in-Q) (IEEE 2006), provider backbone bridge (PBB, also known as MAC-in-MAC) (IEEE 2008), provider backbone bridge traffic engineering (PBB-TE) (IEEE 2009), shortest path bridging (SPB) (IEEE 2012a), transparent interconnection of lots of links (TRILL) (Eastlake et al. 2011), etc. Among them, TRILL and SPB are the most popular techniques appropriate for DCN. TRILL is an Internet engineering task force (IETF) standard for layer-2 scalability. The frame of TRILL protocol introduces a TRILL header between the "outer" and "inner" Ethernet headers. The main fields in the TRILL headers are hop count and ingress and egress nicknames that are used for routing. TRILL network implements layer-2 routing by extending intermediate system to intermediate system (IS-IS) routing protocol (Eastlake et al. 2011), which is similar to the open shortest path first (OSPF) protocol in a layer-3 network. On a TRILL network, data packets can be forwarded using equal-cost multipath (ECMP) algorithm or along the shortest path, which can greatly improve the data forwarding efficiency and throughput of DCNs. Moreover, TRILL can prevent loops and speed up network convergence

in the layer-2 networking. To address the multi-tenancy issue, TRILL uses a 24-bit-long fine-grained label to identify tenants. For SPB, its core idea is similar to TRILL, which also use IS-IS protocol but with different packet encapsulation.

On the other hand, the L2oL3 techniques are also known as network virtualization over layer 3 (NVO3) techniques (Black et al. 2016). NVO3 presents a high-level overview architectural framework that identifies key functional components for data center network virtualization. Currently, four NVO3 techniques are the most popular and supported by the majority of vendors, i.e., virtual extensible local area network (VXLAN) (Mahalingam et al. 2014), network virtualization using generic routing encapsulation (NVGRE) (Garg and Wang 2015), stateless transport tunneling (STT) (Davie and Gross 2016), and generic network virtualization encapsulation (Geneve) (Gross et al. 2018).

VXLAN is an IETF specification that is designed to address the need for overlay networks within virtualized data centers accommodating multi-tenant environment (Mahalingam et al. 2014). VXLAN uses a MAC in user datagram protocol (UDP) encapsulation to enable L2oL3 overlay networks. Each overlay is termed as VXLAN segment and identified through a 24-bit VXLAN network identifier (VNI) specified in VXLAN header. Only VMs using the same VNI can communicate with each other, which can achieve the network isolation for multi-tenant environment. Compared to VLAN, VXLAN allows a larger number of virtual networks (up to 16 M VXLAN segments) to coexist within the same administrative domain. Due to the encapsulation, VXLAN is also considered as a tunneling scheme. The end point of the tunnel termed VXLAN tunnel end point (VTEP) is typically located within the hypervisor on the physical machine that hosts VMs. The VXLAN tunnels (i.e., outer header encapsulation) are known only to the VTEP, while they are transparent to VMs. Thus, it is required that VTEPs should know the association between VM's MAC and VTEP's IP address. To obtain this association, different types of control

schemes can be applied, e.g., either a learning-based control plane or a central authority-based lookup. The central authority-based method is commonly realized by software-defined network (SDN) technique, which will be discussed later.

Similar to VXLAN, both NVGRE and STT are competing specifications but using different encapsulated formats and terms. For example, NVGRE applies a MAC in IP encapsulation, while STT adopts a MAC in transmission control protocol (TCP) encapsulation. Since all three encapsulation methods mentioned above are proposed by different vendors, none of them are compatible with each other. Thus, Geneve is developed by IETF NVO3 working group to address the perceived limitations of the earlier specifications and support all of the capabilities of VXLAN, NVGRE, and STT.

In addition to its advantages, data center network virtualization also brings many other management issues, e.g., configuring and monitoring the huge number of virtual networks for different tenants. Thus, network management technologies are required to efficiently administer and manage network complexity. Network management is not a new topic for data center. In traditional data center, management functions typically reside outside the network in a management system, which interact with network elements via network protocols for management, such as simple network management protocol (SNMP). However, traditional network management systems can only provide limited functionality and flexibility.

Thus, the need for a programmable network has led to much interest in software-defined networking (SDN) technology, which emerges as a new paradigm that shifts network control plane from distributed protocols to a logically centralized controller. SDN helps in network virtualization. Benefited from SDN, flexible network management and rapid deployment of new functionalities can be easily achieved, which makes SDN widely used in data center. The SDN architecture are divided into three planes from top down, i.e., application plane, control plane, and data plan (SDN 2013). In the upper plane, SDN applications define the network requirements and control logic for each tenant, which are sent

to the SDN controller via northbound interface (NBI). As the core entity, the SDN controller is responsible for translating the requirements of SDN applications down to SDN switches and providing the SDN applications with an abstract view of the network. The interface between an SDN controller and an SDN switch is termed control to data-plane interface (CDPI), which provides programmatic control of all forwarding operations, capabilities advertisement, statistics reporting, and event notification. There are various network protocols that can be applied on CDPI, such as SNMP, network configuration protocol (NETCONF) (Enns et al. 2011), open vSwitch database management protocol (OVSDB) (Pfaff and Davie 2013), OpenFlow (OpenFlow 2015), OF-CONFIG (OpenFlow 2014), RESTCONF (Bierman et al. 2017), etc.

Key Applications

There are many key applications including but not limited to multi-tenant data center, data center with a large number of virtual networks, and virtual data center.

Cross-References

- [Data Center Network Virtualization](#)
- [Software-Defined Wireless Networking](#)

References

- (2006) IEEE Standard for Local and Metropolitan Area Networks—Virtual Bridged Local Area Networks—Amendment 4: Provider Bridges. IEEE Std 8021ad-2005 (Amendment to IEEE Std 8021Q-2005) pp 1–74
- (2008) IEEE Standard for Local and metropolitan area networks – Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges. IEEE Std 8021ah-2008 (Amendment to IEEE Std 8021Q-2005) pp 1–110
- (2009) IEEE Standard for Local and metropolitan area networks-Virtual Bridged Local Area Networks Amendment 10: Provider Backbone Bridge Traffic Engineering. IEEE Std 8021Qay-2009 (Amendment to IEEE Std 8021Q-2005) pp 1–131

(2012a) IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—Amendment 20: Shortest Path Bridging

(2012b) IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—Amendment 21: Edge Virtual Bridging. IEEE Std 8021Qbg-2012 (Amendment to IEEE Std 8021Q-2011), pp 1–191

(2013) SDN Architecture Overview. Open Networking Foundation Version 1.0

(2014) IEEE standard for local and metropolitan area networks—bridges and bridged networks. IEEE Std 8021Q-2014 (Revision of IEEE Std 8021Q-2011), pp 1–1832

(2014) OpenFlow Management and Configuration Protocol. Open Networking Foundation Version 1.2

(2015) OpenFlow Switch Specification. Open Networking Foundation Version 1.5.1

Alshaer H (2015) An overview of network virtualization and cloud network as a service. *Int J Netw Manag* 25(1):1–30

Bari MF, Boutaba R, Esteves R, Granville LZ, Podlesny M, Rabbani MG, Zhang Q, Zhani MF (2013) Data center network virtualization: a survey. *IEEE Commun Surveys Tut* 15(2):909–928

Bierman A, Bjorklund M, Watsen K (2017) RESTCONF protocol. <https://tools.ietf.org/html/rfc8040>

Black D, Hudson J, Kreeger L, Lasserre M, Narten T (2016) An architecture for data-center Network Virtualization over Layer 3 (NVO3). <https://tools.ietf.org/html/rfc8014>

Davie B, Gross J (2016) A Stateless Transport Tunneling Protocol for Network Virtualization (STT). <https://tools.ietf.org/html/draft-davie-stt-08>

Eastlake D, Banerjee A, Dutt D, Perlman R, Ghanwani A (2011) Transparent Interconnection of Lots of Links (TRILL) use of IS-IS. <https://tools.ietf.org/html/rfc6326>

Enns R, Bjorklund M, Schoenwaelder J, Bierman A (2011) Network Configuration Protocol (NETCONF). <https://tools.ietf.org/html/rfc6241>

Garg P, Wang P (2015) NVGRE: Network Virtualization Using Generic Routing Encapsulation. <https://tools.ietf.org/html/rfc7637>

Gross J, Ganga I, Sridhar T (2018) Geneve: Generic Network Virtualization Encapsulation. <https://tools.ietf.org/html/draft-ietf-nvo3-geneve-07>

Jain R, Paul S (2013) Network virtualization and software defined networking for cloud computing: a survey. *IEEE Commun Mag* 51(11):24–31

Mahalingam M, Dutt D, Duda K, Agarwal P, Kreeger L, Sridhar T, Bursell M, Wright C (2014) Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks. <https://tools.ietf.org/html/rfc7348>

Pfaff B, Davie B (2013) The Open vSwitch Database Management Protocol. <https://tools.ietf.org/html/rfc7047>

Tatsuhiro A, Osamu S, Katsuhito A (2013) Network virtualization for large-scale data centers. *Fujitsu Sci Tech J* 49(3):8

Wang B, Qi Z, Ma R, Guan H, Vasilakos AV (2015) A survey on data center networking for cloud computing. *Comput Netw* 91:528–547

Xia W, Zhao P, Wen Y, Xie H (2017) A Survey on Data Center Networking (DCN): infrastructure and operations. *IEEE Commun Surv Tutor* 19(1): 640–656

Data Center Networking (DCN): Infrastructure and Operation

Rongtao Xu¹ and Shiwen Liu²

¹State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

²Intelligent Computing and Communication Lab, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Synonyms

[Infrastructure design of data centers](#); [Operation and maintenance of data centers](#)

Definitions

Due to the unique requirements in data center networks (DCNs) such as large scale, high application diversity, high power consumption, strict reliability, etc., the design of infrastructure and operations in DCNs face significant technical challenges. Researchers have been working on respective areas including network topology, optical networking, wireless networking, network virtualization, etc.

Historical Background

The rapid expansion of Internet services make data centers (DC) become a key infrastructure

to promote this growing trend. A DC usually accommodates a large number of computation and storage nodes which are interconnected through a DC network (DCN). As the vital part of communication, DCN plays a key role in DC operation.

Foundations

DCN Infrastructure

Transmission Technology

The data transmission in a DCN can be implemented in a wired or wireless form. Most conventional DCNs are built with electrical cables or optical fiber. In recent years, researchers also work on incorporating wireless technologies in DCNs.

- **Wired DCNs:** For wired DCNs, electrical cables are most commonly used in data centers due to the advantages of convenient deployment and compatibility to the conventional network system. Twisted pair cables can provide transmission rate between 10 Mbps and 10 Gbps with different specifications. However, they often encounter the problems of length limit. In addition to electrical cables, optical fiber is also widely used in data centers. Optical fiber has higher bandwidth, smaller size, lighter weight, and lower power consumption, compared to electrical cables. Using techniques like WDM (wavelength division multiplexing), 100 Gbps transmission rate in optical links has been achieved (Jianjun et al. 2016).
- **Wireless DCNs:** Due to the limitations and problems in the existing wired DCNs, researchers have been working on incorporating wireless communication technologies in DCNs. Up to now, 60 GHz radio frequency (RF) and free space optics (FSO) are two candidate technologies which are mostly discussed.

Networking Topologies

In switch-centric topologies, switches are mainly responsible for network construction and data transmission. The switches are usually connected by hierarchy topologies and the servers are generally connected to the low-level switches at network edge. The most typical topology in wired DCN is the tree-based hierarchical topology.

In fat-tree topology, all switching elements are identical, which provides cheap commodity components for all switches in communication architecture. In addition, for any communication mode, there are some paths that will saturate all the bandwidth available to the terminal host in the topology. In practice, it may be difficult to achieve an oversubscription ratio of 1:1 because of the need to prevent packet reordering of TCP flows (Al-Fares et al. 2008).

VL2 is a practical network architecture that can be extended to support large data centers with uniform high capacity between servers, performance isolation between services, and Ethernet layer semantics. VL2 uses flat addressing to allow service instances to be placed anywhere in the network. Address resolution based on terminal system can be extended to large server pools without complication to the network control plane (Greenberg et al. 2009).

In Walraed-Sullivan et al. (2013), the author uses the Aspen tree as a technique to modify the hierarchical data center topology, enabling the switch to respond locally to faults, which reduces the convergence time and control overhead of fault recovery. Studies have found that reducing the convergence time of a topology for a given network size results in a proportional decrease to its scalability. On the other hand, reducing the convergence time without affecting scalability requires the introduction of additional switches and links.

In a server-centric topology, servers are used to compute, route, or relay data to each other. DCell is a network structure with many functions for data center network. It is a recursive structure, in which advanced DCells are made up of many low-level DCells. The DCell at the same level is completely connected to each other.

DCell is fault-tolerant because it has no single point of failure, and its distributed fault-tolerant routing protocol will execute routes close to the shortest path even if there are serious link or node failures. For all types of services, it also provides higher network capacity than traditional tree-based structures (Guo et al. 2008).

DCN Operations

Traffic Control

The path to forward packets in the DCN is determined by various protocols, such as spanning tree, VLAN, routing algorithm, and OpenFlow. In a data center network, routers need to go beyond Layer 2 and the interconnected network with different network addresses. There are also many routing algorithms designed for specific network topologies, such as DCell Fault Tolerant Routing for DCell, BCube Source Routing Protocol for BCube, and Traffic Aware Routing for Traffic Conn.

Data center services also require priority management. Priority management provides different qualities of service by processing packets according to packet-based priorities rather than arrival orders. When a packet arrives at a transmitting or receiving queue with a full buffer, it is discarded if its priority is less than or equal to the lowest priority. Otherwise, the packet with the lowest priority is discarded to make room for the new packet. When the port is idle, the packet with the highest priority is sent.

Service Latency

Some service in data centers face rigorous service latency requirement. D3 (deadline-driven delivery) in Wilson et al. (2011) introduces a deadline-aware rating allocation for flows. Flows with deadlines encode the desired rate in their TCP SYN packets. Inspired by the recommendation of managing network congestion through explicit rate control, D3 explores the feasibility of using deadline information to control the rate at which terminal hosts introduce traffic in the network. Specifically, applications open process expiration date and size information when the process starts. The terminal host uses this information to request

the speed of the router from the data path to the destination.

Network Maintenance

The maintenance of the DCN is required after deployment, in order to monitor the network, identify network failures, and carry out repair actions. Techniques for detecting and handling network failures are studied by researchers. In Wu et al. (2012), it is recommended that NetPilot automatically mitigates network failures rather than repair faults. NetPilot mitigates failures in the same way as the operator – by deactivating or restarting suspected violation components. NetPilot does not need to know the exact root cause of the fault. Instead, it is implemented by an intelligent trial-and-error. The core includes an Impact Estimator, which prevents excessive disruptive mitigation measures, and a fault-specific mitigation planner that minimizes the number of trials.

Key Applications

Data center networking techniques can be used in data centers to ensure the performance of data centers.

Cross-References

- [Data Center Network Virtualization](#)
- [Wireless Network](#)

References

- Al-Fares M, Loukissas A, Vahdat A (2008) A scalable, commodity data center network architecture, pp 63–74
- Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S (2009) V12: a scalable and flexible data center network. *Commun ACM* 54(4):95–104
- Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S (2008) Dcell: a scalable and fault-tolerant network structure for data centers. *ACM SIGCOMM Comput Commun Rev* 38(4):75–86

- Jianjun Y, Junwen Z (2016) Recent progress on high-speed optical transmission. *Digit Commun Netw* 2(2):65–76
- Walraed-Sullivan M, Vahdat A, Marzullo K (2013) Aspen trees: balancing data center fault tolerance, scalability and cost. In: *ACM conference on emerging NETWORKING experiments and technologies*, pp 85–96
- Wilson C, Ballani H, Karagiannis T, Rowtron A (2011) Better never than late: meeting deadlines in datacenter networks. *ACM SIGCOMM Comput Commun Rev* 41(4):50–61
- Wu X, Turner D, Chen CC, Maltz DA, Yang X, Yuan L, Zhang M (2012) Netpilot: automating datacenter network failure mitigation. In: *ACM SIGCOMM 2012 conference on applications, technologies, architectures, and protocols for computer communication*, pp 419–430

Data Collection

- [Data Gathering in Wireless Sensor Networks](#)

Data Dissemination

- [Data Dissemination in Delay Tolerant Networks with Geographic Information](#)

Data Dissemination in Delay Tolerant Networks with Geographic Information

Longxiang Gao¹, Tom H. Luan², Shui Yu³, and Wanlei Zhou³

¹Deakin University, Burwood, VIC, Australia

²Xidian University, Xi'an, China

³University of Technology Sydney, Ultimo, NSW, Australia

Synonyms

[Data dissemination](#); [Data forwarding](#); [Delay tolerant networks](#); [Opportunistic networks](#)

Definitions

The delay-tolerant networks (DTNs) are under the umbrella of mobile ad hoc networks, which are featured with the long latency and unstable wireless topology. In DTNs, the end-to-end transmission delay from one node to another can be measured in days due to the absence of the routing paths.

Data forwarding in DTNs, such as data ferry, typically adopts store-and-forward techniques: if there are no connections available at a particular time, the source node needs to store and carry the message until the next available node is encountered, and these encountered nodes further relay the contents until the destination in the same store-and-forward manner.

Historical Background

Traditional data forwarding protocols in mobile ad hoc networks include two categories: proactive and reactive routings. In proactive routings, forwarding routes are computed automatically and independent of traffic arrivals. Examples of proactive routings include the Destination Sequenced Distance Vector (DSDV) (Perkins and Bhagwat 1994) and Optimized Link State Routing (OLSR) (Dhurandher et al. 2010). These protocols are capable of computing routes for a connected graph. However, they cannot find a path to link nodes in a disconnected but dynamic graph, such as DTNs.

In reactive routings, routes are discovered on demand when traffic needs to be delivered to an unknown destination. Such routing protocols include the Ad-hoc On-demand Distance Vector (AODV) (Perkins and Royer 1999) and Dynamic Source Routing (DSR) (Wu 2002). In these systems, a route discovery technique is employed to determine a path to the destination. These on-demand routing protocols incur additional delay, and they work best when communication patterns are relatively sparse. In delay-tolerant networks, as with proactive protocols, these protocols only work for finding routes in a connected subgraph

of the overall DTNs routing graph. However, they fail differently to proactive protocols. In particular, they simply fail to return a successful route (from a lack of response), whereas proactive protocols can potentially fail quicker (by determining the requested destination is not presently reachable).

The data forwarding issue is the fundamental problem for delay-tolerant networks. To deliver messages in such networks, existing algorithms, such as Epidemic (Lindgren et al. 2003) and PRoPHET (Lindgren et al. 2003), use flooding or partial flooding with a probability. There is a high probability that any flooding may cause network congestion or high interference, which consumes further resources for processing and switching operations. Context-aware routing protocols, such as CAR (Musolesi and Mascolo 2009) and HiBOP (HiB 2007), utilize context information such as history, battery status, and changes to the rate of connectivity to compute delivery probabilities. One of the reasons for the above routing mechanisms to use flooding or partial flooding is the lack of routing information. Geographic characters are unreliable and not enough to route messages efficiently due to the nodes' mobility characteristics. However, given the dynamic nature of DTNs, it should take other characteristics into consideration as a viable and beneficial option when we look into ways of increasing efficiency.

Foundations

Vector Routing Protocol

Kang and Kim (2008) propose a vector routing scheme in delay-tolerant networks in which the nodes own location information to calculate their velocity and direction to destination, which reduces the overhead of routing without decreasing the delivery success ratio. In Kang and Kim (2008), a mobile node uses its current vector V_{cur} and history movement pattern to predict a future moving vector. If two nodes are moving in the same direction based on the moving vector, the replication of packets for both nodes is not efficient. On the other hand, if two nodes move

in different directions, it is necessary to forward packets to both nodes. In vector routing, nodes moving in an orthogonal direction have a higher packet replication frequency compared to nodes moving in similar or opposite directions.

Furthermore, the velocity factor has also been taken into consideration. If two nodes move in the same direction, but the first node is faster than the second, packets may be forwarded to the faster node instead of replicating the slower node again. Simulations conducted from the NS-2 platform demonstrated that vector routing has a similar delivery success ratio compared with epidemic routing, but with 28% less traffic than random way-point model and 38% less traffic than Manhattan mobility model.

Similarity-Based Mobility Pattern-Aware Routing Protocol

Yin et al. (2009) propose a similarity-based mobility pattern-aware routing in DNTs, which uses GPS to detect the location of nodes and calculates their mobility pattern, where the more identical the mobility pattern is, the more stable their relative positions are. The mobility pattern used in Yin et al. (2009) is the intermediate node's (node r) speed direction to the destination node (node d), which is an angle (θ_{rd}).

Based on the nodes location information ((x_r, y_r) and (x_d, y_d)) and the weight controller (λ_i), a similarity degree ($f(r, d)$) is proposed to select the next forwarding node with the highest similarity degree in the source node's one-hop neighbor area

$$f(r, d) = \lambda_1 \times \frac{1}{\theta_{rd}} + \lambda_2 \times \frac{1}{\sqrt{(x_r - x_d)^2 + (y_r - y_d)^2}}. \quad (1)$$

Location-Aware Routing Protocol

Tian and Li (2010) propose a location-aware routing protocol to study the node isolation issue in DTNs. In this routing protocol, analysis of both the location information and social network structure is conducted. When the current node and the destination node are within the same

component and if the encountered node has a bigger metric than the current node, the packet is forwarded to this encountered node. Otherwise, the current node just wait next available nodes.

Furthermore, Tian and Li (2010) differentiates the operations of forwarding and replicating from existing literature. In Tian and Li (2010), a node can forward a message to an encounter node once only, and then it becomes forwarding inactive, even though it can replicate a message to other nodes without such constraint. Therefore, when the current node is in a different component to the destination node and if the current node has no previous contact with the encountered node, the packet is replicated to the encountered node. Otherwise, the packet is forwarded to the encountered node.

Location-Aided Routing Protocol

Ko and Vaidya (2000) propose location-aided routing (LAR) to show how to use the location information to reduce routing overhead by limiting the search space for a certain route. The traditional routing discovery is based on the flooding process, where source nodes broadcast route request messages. When an internode receives this request message, it compares the message destination with its own identifier. If they match, this message arrives at the final destination; otherwise, the internode needs to broadcast this message again.

LAR uses location information to reduce the number of propagated nodes where they define two concepts: the expected zone and the request zone. The expected zone is a location where a certain node has visited before. With the expected zone in mind, the forwarding process can bypass lots of useless internodes. If the source node has no idea where the expected zone is, the entire region then becomes the expected zone, and the flooding process is deployed. The request zone defines a route requesting area (flooding area), which includes the expected zone and its surrounding zones. Therefore, the size of the request zone determines the routing performance and the network overhead. One potential method to determine the size of the request zone based LAR is to find the smallest rectangle that includes

the current location and the expected zone. This expected zone is a circle area, where its center is the previous location of the destination node and its radius is the product of the time period and the node's movement speed. Therefore, the size of this request zone is subject to the average node's movement speed and the time elapsed since the last recorded known location.

Geography-Aware Active Routing Protocol

Fan et al. (2010) propose a geography-aware active data dissemination protocol in human-associated DTNs that analyzes mobile social networks from a geographic point of view. This paper uses the "community" concept from social science, which is defined as a clustering of users closely linked to each other and extends it to "geo-community" by considering the relationship of nodes on geographic related social networks. The higher the centrality of a geo-community, the better capable it is of contacting mobile users in social networks. To quantify the dynamic density of a geo-community, the following geo-centrality technique is applied as:

$$C_i(t_i) = 1 - \frac{1}{N} \sum_{k=1}^N (1 - \phi_i^k)^{t_i}, \quad (2)$$

where $C_i(t_i)$ indicates the average probability a certain user visits geo-community i at time t_i and ϕ_i^k is used to measure the steady-state probability distribution for node k in its i th geo-community. With this geo-community, the data forwarding process can be conducted by combing social and geographic information.

Key Applications

There are many applications based on delay-tolerant networks. Some typical applications can be classified as digital communication for rural areas and personal/wildlife communications.

Applications for digital communication at rural areas include:

- DakNet (Pentland et al. 2004) uses existing communications and transport infrastructure to form a delay-tolerant network, where it combines physical transportation with a wireless data transferral function to extend Internet connectivity to a central hub, such as cyberspace or a post office. DakNet does not need to relay data over a long distance, which is an advantage because it would reduce costs and save a lot of power.
- TrainNet (Zarafshan-Araki and Chin 2010) provides a low-cost and high-bandwidth link. The reason to use the train instead of other transportation methods is its fixed timetable where movements are deterministic and cover long distance. It can also carry more storage devices to store more data. The potential application of TrainNet intends to deliver a large non real-time video among any city as long as it has a train system.
- KioskNet (Seth et al. 2006) extends the idea from DakNet (Pentland et al. 2004) to have a complete system with naming, addressing, forwarding, and routing. In this system, a kiosk controller is used to provide network boot, storage, and user management functions. For those public users, they use the kiosk controller to access the executed applications. For organization users, such as government officials and NGO partners, they treat the kiosk controller as a wireless hotspot to provide a DTN function (store-and-forward) to the Internet. The main advantage of this work is the use of mechanical backhaul as ferries to carry data to the Internet gateway as they pass a kiosk. The mechanical backhaul can be cars, buses, motorcycles, or trains. During a one-way communication period (20 s to 5 mins), a 100 KB to 50 MB data can be transferred.

Applications for personal/wildlife communications are:

- Pollen network (Glance et al. 2001) borrows the idea that a plant is pollinated by insects.

For example, when an insect visits a flower to take nectar, it needs to move to a different flower as some pollen rubs off. Using the same method, when a mobile device carrier passes by an artifact, he or she can leave comments on the associated device, and it can also transfer pieces of pollen to other mobile devices. However, Pollen is not suitable for applications requiring a fast response or guaranteed delivery. Instead, it provides communication between large numbers of objects and devices, which are too expensive to be networked with the current infrastructure.

- Wireless Body Area Network (WBAN) (Qiao et al. 2011; Jovanov et al. 2005) solves on body packet routing issues in the presence of topological partitioning. Sensor nodes in WBAN are based on low-power supplied RF transceivers (Qiao et al. 2015; Cheng et al. 2014) where the signal transmission can be affected by postural body movements and clothing.
- ZebraNet (Juang et al. 2002) is designed to track wildlife in Kenya. Costumed tracking collars carried by animals are used to collect a zebra's mobility pattern and send the pattern back to a mobile base station. As the network between zebras and base stations is opportunistic, the DTN technique is used for this system. By using the DTN technique, ZebraNet stores and forwards collected mobility patterns and receives updated information from the mobile base station, which may not be available at a particular time. In addition, when base stations receive data, there may be no Internet connection, and the data has to remain at the base station.

References

- Boldrini C, Conti M, Jacopini J, Passarella A (2007) Hibop: a history based routing protocol for opportunistic networks. In: Proceedings of IEEE WoW-MoM, pp. 1–12. <https://doi.org/10.1109/WOWMOM.2007.4351716>

- Cheng M, Ye Q, Cai L (2014) Rate-adaptive concurrent transmission scheduling schemes for wpans with directional antennas. *IEEE Trans Veh Technol*. <https://doi.org/10.1109/TVT.2014.2365496>
- Dhurandher S, Obaidat M, Gupta M (2010) A reactive optimized link state routing protocol for mobile ad hoc networks. In: *Proceedings of IEEE ICECS*, pp 367–370. <https://doi.org/10.1109/ICECS.2010.5724529>
- Fan J, Du Y, Gao W, Chen J, Sun Y (2010) Geography-aware active data dissemination in mobile social networks. In: *Proceedings of IEEE MASS*, pp 109–118. <https://doi.org/10.1109/MASS.2010.5663960>
- Glance N, Snowdon D, Meunier JL (2001) Pollen: using people as a communication medium. *Comput Netw* 35(4):429–442. [https://doi.org/10.1016/s1389-1286\(00\)00183-3](https://doi.org/10.1016/s1389-1286(00)00183-3)
- Jovanov E, Milenkovic A, Otto C, de Groen PC (2005) A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. *J Neuroeng Rehabil* 2(1):6. <https://doi.org/10.1186/1743-0003-2-6>
- Juang P, Oki H, Wang Y, Martonosi M, Peh LS, Rubenstein D (2002) Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet. *SIGPLAN Not* 37(10):96–107. <https://doi.org/10.1145/605432.605408>
- Kang H, Kim D (2008) Vector routing for delay tolerant networks. In: *Proceedings of IEEE VTC*, pp 1–5. <https://doi.org/10.1109/VETECF.2008.21>
- Ko YB, Vaidya NH (2000) Location-aided routing (LAR) in mobile ad hoc networks. *Wirel Netw* 6(4):307–321. <https://doi.org/10.1023/A:1019106118419>
- Lindgren A, Doria A, Schelén O (2003) Probabilistic routing in intermittently connected networks. *SIGMOBILE Mob Comput Commun Rev* 7(3):19–20. <https://doi.org/10.1145/961268.961272>
- Musolesi M, Mascolo C (2009) CAR: context-aware adaptive routing for delay-tolerant mobile networks. *IEEE Trans Mob Comput* 8(2):246–260
- Pentland A, Fletcher R, Hasson A (2004) DakNet: rethinking connectivity in developing nations. *Computer* 37(1):78–83. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1260729
- Perkins CE, Bhagwat P (1994) Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. In: *Proceedings of the conference on communications architectures, protocols and applications, SIGCOMM'94*. ACM, New York, pp 234–244
- Perkins C, Royer E (1999) Ad-hoc on-demand distance vector routing. In: *Proceedings of IEEE WMCSA*, pp 90–100
- Qiao J, Cai LX, Shen X, Mark JW (2011) Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks. *IEEE Trans Wirel Commun* 10(11):3824–3833
- Qiao J, Shen XS, Mark JW, Shen Q, He Y, Lei L (2015) Enabling device-to-device communications in millimeter-wave 5G cellular networks. *IEEE Commun Mag* 53(1):209–215
- Seth A, Kroeker D, Zaharia M, Guo S, Keshav S (2006) Low-cost communication for rural internet kiosks using mechanical backhaul. In: *Proceedings of the 12th annual international conference on Mobile computing and networking, MobiCom'06*. ACM, New York, pp 334–345. <https://doi.org/10.1145/1161089.1161127>
- Tian Y, Li J (2010) Location-aware routing for delay tolerant networks. In: *Proceedings of ICST CHINACOM*, pp 1–5
- Wu J (2002) An extended dynamic source routing scheme in ad hoc wireless networks. In: *Proceedings of annual Hawaii international conference on system sciences*, pp 3832–3838. <https://doi.org/10.1109/HICSS.2002.994517>
- Yin L, Cao Y, He W (2009) Similarity degree-based mobility pattern aware routing in DTNs. In: *Proceedings of the 2009 international symposium on intelligent ubiquitous computing and education, IUCE'09*. IEEE Computer Society, Washington, DC, pp 345–348. <https://doi.org/10.1109/IUCE.2009.84>
- Zarafshan-Araki M, Chin KW (2010) TrainNet: a transport system for delivering non real-time data. *Comput Commun* 33(15):1850–1863. <https://doi.org/10.1016/j.comcom.2010.06.008>

Data Falsification in Wireless Data Fusion

- Byzantine Attack and Defense in Wireless Data Fusion

Data Forwarding

- Data Dissemination in Delay Tolerant Networks with Geographic Information

Data Gathering and Aggregation with Privacy Preservation in Smart Grids

- Privacy-Preserving Data Aggregation for Smart Grids

Data Gathering in Wireless Sensor Networks

Catalina Aranzazu-Suescun and Mihaela Cardei
Florida Atlantic University, Boca Raton, FL,
USA

- *Quality of service.* Various applications have different quality-of-service requirements, such as data delay, data freshness, and reliable data delivery. For example, some applications may be delay tolerant, while other applications have stringent delay requirements.

Synonyms

Data collection; Event detection and reporting

Definitions

Data gathering is an important operation in wireless sensor networks which deals with collecting data from the sensor nodes and transmitting them to a sink (or base station) where it is further processed and available for end-user queries.

Historical Background

Wireless sensor networks (WSNs) are deployed to collect useful data from an area of interest. Data gathering is an important operation in wireless sensor networks, and it deals with collecting data from the sensor nodes. There are several aspects that should be considered when designing algorithms and protocols for data gathering:

- *Energy efficiency.* Sensor nodes are usually resource constrained, and they are equipped with batteries which often cannot be recharged or replaced. Therefore, it is important to design energy-efficient mechanisms.
- *Network elements.* There are several variations in the architecture of the network, such as the following: sensor nodes may be homogeneous or heterogeneous (e.g., they may be equipped with different types of sensing components), some of the nodes may be resource-rich devices which can serve as cluster heads or gateways; node distribution in the field; and the sink can be fix or mobile.

Sensors may be pre-programmed to measure certain parameters, or a sink may inject a request (or interest) into the network. There are few types of data delivery which trigger the data gathering protocol: periodic data delivery, query data delivery, and event-based data delivery. When the sink is responsible to inject the request or query into the network, it broadcasts a message with the requirements of the event to be monitored, such as the location, time, and threshold for the environmental measurements. Sensor nodes which receive the broadcast message and satisfy the location and time requirements, task their sensors to collect samples, and if the measurements exceed the thresholds, then they send the report to the sink.

In the periodic data delivery model, the nodes send data reports to the sink periodically, at regular time intervals, even if no specific event is present in the area. When the data delivery is event-based, the nodes only send information to the sink when an event is detected in the monitoring area. Finally, when the data delivery is query-based, the nodes in the network send information to the sink based on the requirement of the sink.

Nodes can deliver the data report to the sink directly, or using a certain infrastructure. For example, the data can be sent using a converge-cast tree rooted at the sink. One issue that can occur in this case is an *energy hole* around the sink, where the nodes close to the sink deplete their energy resources at a faster rate and eventually die, resulting in network partitioning. One option to deal with this issue is to use a certain infrastructure, such as clustering. Each cluster has a cluster head (CH) which is usually responsible for collecting and aggregating data from its cluster.

The process of finding a route to the sink can be done using a route request query (RREQ) or using a tree. In the RREQ approach, if a node has data to send to the sink, then it broadcasts a RREQ message in the network. A node that receives the RREQ message and knows a route to the sink sends back a route reply (RREP) message containing information about its route to the sink. In a tree-based approach, when the sink floods the network with an event request, a node receiving the request message will set up as its parent the node from which the message was received. In this way, a convergecast tree is formed, and assuming that the network is connected, each node has a path to the sink.

Data aggregation is an important operation in wireless sensor networks, which interleaves with data gathering. The objective is to combine data from different source nodes en route, thus eliminating redundancy and reducing the number of packets in the network, which results in energy saving. For cluster-based mechanisms, a natural approach is to have the CH perform the aggregation of the data messages in the cluster.

The clustering mechanism can be infrastructure-based or event-based. In the infrastructure-based clustering, the area is partitioned into fixed cells, and each cell has a CH. One approach is to have the node closest to the center of the cell assume the role of CH. In the event-based approach, clusters are formed on demand, only when an event is detected in the monitoring area.

Foundations

There are several design choices which affect data gathering. Next, we discuss some of the most popular architectures proposed in literature.

Homogeneous WSN with Static Sink

Direct diffusion (Intanagonwiwat et al. 2000) is an important data gathering mechanism. The sink injects a query, called interest, in the sensor network specifying the event to be monitored. This interest is diffused in the whole network. Sensor nodes located in the region of interest task their sensing components to start collecting

the information. As the interest propagates in the network, nodes record gradients which allow them to send data back to the sink. Once the sink starts receiving data, it may decide to reinforce some of the representative events. Samaras and Triantari (2016) use clustering to improve the direct diffusion mechanism. Clusters are formed with nodes that have the same interests, and the CH is selected based on the node energy level, distance to the sink, and network topology.

Heterogeneous WSN with Static Sink

In a heterogeneous network, the nodes have different sensing components or different capabilities. For example, a network may contain several resource-rich devices that can be used as cluster heads or gateways. Composite events (Gao et al. 2015) can be used to model complex events, such as fire, when more sensing attributes are needed to accurately detect the event. A composite event can be defined using one or more atomic events, where each atomic event is triggered when a single sensing value (or attribute) exceeds some threshold. In a cluster-based network model, the CHs are usually responsible for sending data to the sink. Clusters could be formed since the deployment of the network (Memon and Muntean 2012; Khan et al. 2015), or they could be triggered by the detection of an event. Event-based clustering mechanisms are proposed in the works of Aranzazu-Suescun and Cardei (2016) and Villas et al. (2012). An event-based clustering mechanism usually leads to a smaller number of clusters compared to fixed (or predefined) clustering. A smaller number of clusters detecting and reporting the events may result in energy savings. However, the process of forming the event-based clustering may add some delays in data reporting.

WSNs with Multiple Sinks

Marta and Cardei (2009) study data gathering in WSNs with multiple static sinks. The area is divided into hexagonal tiling, and sinks are located in the center of each tile. Each sink locally collects data. The study shows that the nodes closer to the sink consume more energy and are susceptible to the energy hole problem. The paper

argues that using mobile sinks can alleviate this issue.

WSNs with Mobile Sink (or Sinks)

Some of the models used for mobile sinks are random walk movement, predefined path, and controlled path. When a random walk movement of the sink is employed, the data delivery paths may become obsolete as a result of sink movement. An important concept used to deal with sink movement is the use of anchor nodes. An anchor node is usually located close to the sink, and it acts as a relay in sending data to the sink. In the articles of Tunca et al. (2015) and Hu et al. (2015), the sink selects a new anchor every time it loses its connectivity with the current anchor, as a result of moving. After a new anchor is selected, usually some information about the new anchor is flooded into the network, so that the nodes in the network revise their path to the sink. To avoid frequent updates in the network which waste energy, some authors have used multiple anchors. In this case, a new anchor is selected when the sink loses connectivity to the old anchor. Then, the process is repeated for a determined number of times before the first anchor is reselected and the process is restarted (Aranzazu-Suescun and Cardei 2017; Tian et al. 2010).

If the sink has a predefined path, then the path is usually defined by a number of polling or rendezvous points (RPs). In this case, the sink moves to each polling or rendezvous point and collects information from the closer nodes. Different approaches have been used to select the best paths to visit the RPs. Some authors Salarian et al. (2014) and Ma et al. (2013) have used different mechanisms to solve the traveling salesman problem (TSP). The TSP is NP-complete, and various approximations of the optimal tour have been proposed in literature. In the article of Ma and Yang (2008), the tour of the sink is computed using the polling points as a spanning covering tree.

In some articles, the network is divided into zones, and the path of the sink could be circular, rectangular, or linear depending on the distribution of the regions. In the work of Marta and Cardei (2009), the sinks have predefined

paths, along hexagonal regions. In the article of Abo-Zahhad et al. (2015), the sink computes its sojourn location in each region and the location of the optimal CH using the adaptive immune algorithm.

In the controlled path movement model, the sink movement depends on a specified rule in the network. One option is to trigger the sink movement when the energy of the nearby sensors decrease below some specific threshold (Khan et al. 2012; Marta and Cardei 2009). A fitness equation can be used to determine the neighboring nodes that have the higher residual energy and are closer to the sink (Patel and Bhagat 2014).

Autonomous underwater vehicles (AUVs) can be used in an underwater WSN as collectors (Javaid et al. 2015). AUVs move in the network along a predefined path gathering the information collected by gateway nodes, where the gateway nodes receive data from their neighbor nodes.

Disconnected (Partitioned) WSNs

When the network is partitioned, some mobile entities are used to collect information from different zones in the network. These mobile devices are called MULEs. One such example is using MULEs in the agricultural sensor networks (Burrell et al. 2014) where deploying a dense sensor network can be expensive. Nodes are deployed in some strategical zones, and a MULE gathers the information from each subzone. If the field is too big, then having only one MULE is not suitable because its battery can be quickly drained. The article of Jea et al. (2006) uses multiple MULEs that walk parallel in straight lines in a field. Each node in the field is assigned to a MULE such that the nodes are evenly distributed among the MULEs. A node sends data when it receives a poll message from its assigned MULE. In the work of Shah et al. (2003), the area is divided into a grid, where cells contain sensors or access points. MULEs are used to transfer data from the sensor nodes to access points. The MULEs use a random walk mobility model.

The use of unmanned aerial vehicles (UAVs) in WSNs has increased in recent years. UAVs have been used as mobile sinks and as relay nodes. In the work of Dong et al. (2014), a drone

acts as a mobile sink sending agents to the leaders of the clusters in the network. The agents collect and aggregate the information in the network and then send it to the UAV. UAVs can be used to interconnect disjoint network partitions. UAVs can travel to isolated islands of sensor nodes to collect data. If multiple UAVs are available, they can form a virtual backbone used to communicate with the sensor network (Pignaton de Freitas et al. 2010).

Key Applications

Data gathering is present in all applications of wireless sensor networks, and it should be designed according to applications' quality-of-service requirements and network characteristics.

Cross-References

- [Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks](#)
- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)
- [Information-Centric Wireless Sensor Networks](#)
- [Protocol Stack Architecture for Wireless Sensor Networks](#)

References

- Abo-Zahhad M, Ahmed SM, Sabor N, Sasaki S (2015) Mobile sink-based adaptive immune energy efficient clustering protocol for improving the lifetime and stability period of wireless sensor networks. *IEEE Sensors J*, vol 15. <https://doi.org/10.1109/JSEN.2015.2424296>
- Aranzazu-Suescun C, Cardei M (2016) Event-based clustering for composite event detection in wireless sensors networks. In: 35th IEEE international performance computing and communications conference (IPCCC), Las Vegas
- Aranzazu-Suescun C, Cardei M (2017) Distributed algorithms for event reporting in mobile-sink WSNs for internet of things. *Tsinghua Sci Technol J*, vol 22. <https://doi.org/10.23919/TST.2017.7986944>
- Burrell J, Brooke T, Beck R (2014) Vineyard computing: sensor networks in agricultural production. *IEEE Pervasive Comput J*, vol 3. <https://doi.org/10.1109/MPRV.2004.1269130>
- Dong M, Ota K, Lin M, Tang Z, Du S, Zhu H (2014) UAV-assisted data gathering in wireless sensor networks. *J Supercomput*, vol 70. <https://doi.org/10.1007/s11227-014-1161-6>
- Gao J, Li J, Cai Z, Gao H (2015) Composite event coverage in wireless sensor networks with heterogeneous sensors. In: 2015 IEEE conference on computer communications (INFOCOM), Hong Kong
- Hu YF, Ding LH, Ren LH, Hao KR, Han H (2015) An endocrine cooperative particle swarm optimization algorithm for routing recovery problem of wireless sensor networks with multiple mobile sinks. *Info Sci J*, vol 300. <https://doi.org/10.1016/j.ins.2014.11.052>
- Intanagonwiwat C, Govindan R, Estrin D (2000) Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: The 6th annual international conference on mobile computing and networking, Boston
- Javaid N, Ilyas N, Ahmad A, Alrajeh N, Qasim U, Ali Khan Z, Liaqat T, Iqbal Khan M (2015) An efficient data-gathering routing protocol for underwater wireless sensor networks. *Sens J*, vol 11. <https://doi.org/10.3390/s151129149>
- Jea D, Somasundara A, Srivastava M (2006) Multiple controlled mobile elements (Data Mules) for data collection in sensor networks. In: 1st international conference on distributed computing in sensor systems, San Francisco
- Khan MI, Gansterer WN, Khokhar A, Haring G (2012) Static vs mobile sink: the influence of basic parameters on energy efficiency in wireless sensor networks. *Comput Commun J*, vol 15. <https://doi.org/10.1016/j.comcom.2012.10.010>
- Khan AW, Abdullah AH, Razzaque MA, Bangash JI (2015) VGDR: a virtual grid-based dynamic routes adjustment scheme for mobile sink-based wireless sensor networks. *IEEE Sens J*, vol 36. <https://doi.org/10.1016/j.comcom.2012.04.007>
- Ma M, Yang Y (2008) Data gathering in wireless sensor networks with mobile collectors. In: IEEE international symposium on parallel and distributed processing (IPDPS), Miami
- Ma M, Yang Y, Zhao M (2013) Tour planning for mobile data-gathering mechanisms in wireless sensor networks. *IEEE Trans Veh Technol*, vol 62. <https://doi.org/10.1109/TVT.2012.2229309>
- Marta M, Cardei M (2009) Improved sensor network lifetime with multiple mobile sinks. *Elsevier J Pervasive Mob Comput*, vol 5. <https://doi.org/10.1016/j.pmcj.2009.01.001>
- Memon I, Muntean T (2012) Cluster-based energy-efficient composite event detection for wireless sensor networks. In: The sixth international conference on sensor technologies and applications (SENSORCOMM), Rome
- Patel B, Bhagat D (2014) Shifting of sink position in wireless sensor networks. *Int J Eng Dev Res* 2:2566–2571
- Pignaton de Freitas E, Heimfarth T, Netto IF, Lino CE, Pereira CE, Morado Ferreira A, Wagner FR, Larsson

- T (2010) UAV relay network to support WSN connectivity. In: International congress on ultra-modern telecommunications and control systems and workshops (ICUMT), Moscow
- Salarian H, Chin KW, Naghdy F (2014) An energy-efficient mobile-sink path selection strategy for wireless sensor networks. *IEEE Trans Veh Technol*, vol 63. <https://doi.org/10.1109/TVT.2013.2291811>
- Samaras NS, Triantari FS (2016) On direct diffusion routing for wireless sensor networks. In: 2016 advances in wireless and optical communications (RTUWO), Riga
- Shah RC, Roy S, Jain S, Brunette W (2003) Data MULEs: modeling and analysis of a three-tier architecture for sparse sensor networks. In: 1st IEEE international workshop on sensor network protocols and applications, Anchorage
- Tian K, Zhang B, Huang K (2010) Data gathering protocols for wireless sensor networks with mobile sinks. In: IEEE global telecommunications conference (GLOBE-COM10), Miami
- Tunca C, Isik S, Donmez MY, Ersoy C (2015) Ring routing: an energy-efficient routing protocol for wireless sensor networks with a mobile sink. *IEEE Trans Mob Comput*, vol 14. <https://doi.org/10.1109/TMC.2014.2366776>
- Villas LA, Boukerche A, Guidoni DL, de Oliveira H, Borges de Araujo R, Loureiro A (2012) An energy-aware spatio-temporal correlation mechanism to perform efficient data collection in wireless sensor networks. *Comput Commun J*, vol 36. <https://doi.org/10.1016/j.comcom.2012.04.007>

Data Mining

- [Data-Driven Pattern Prediction](#)

Data Privacy

- [Architecture and Data Management for Smart Community Information Platform](#)

Data Specified Security

- [Data-Driven Security](#)

Data Streams

- [Data Analytics for Longitudinal Biomedical Data](#)

Data Transmission

- [Power Control for Data Transmission in Wireless Networks](#)

Data-Access QoE Management

- [Data-Driven QoE Measurement](#)

Data-Aware QoE Management

- [Data-Driven QoE Measurement](#)

Database

- [Database-Based Spectrum Access](#)

Database-Based Spectrum Access

Haibo Zhou and Xuesen Peng
Nanjing University, Nanjing, China

Synonyms

[Cognitive radio](#); [Database](#); [Dynamic spectrum access](#); [TV white space](#)

Definition

Database-based spectrum access is a popular TV White Space (TVWS) spectrum access manner where the unlicensed users attempt to access the unused or underutilized TV bands based on a database which contains the information related to spectrum availability at a given location and time.

Historical Background

As far back as 2004, known as the first worldwide effort to utilize TVWS, IEEE 802.22 Working Group was formed to define a standard for the license-exempt devices using the TVWS spectrum which used to be occupied by TV broadcast services (Cordeiro et al. 2006). In 2010, as a milestone in TVWS exploitation, Federal Communications Commission (FCC) ruling on TVWS proposed a database-based manner to determine TVWS availability (FCC 2018). In addition, IEEE 802.11af Working Group was formed in 2010 to adopt IEEE 802.11 protocol for TV band operation. IEEE 802.11af is also called super WiFi or WhiteFi, which is designed to leverage the success of WiFi while addressing its limitations (Flores et al. 2013). Different from IEEE 802.22 which fits the outdoor scenario, IEEE 802.11af takes both the outdoor and indoor scenarios into account.

In the next few years, the database-assisted TVWS access manner attracts more and more attentions owing to the supports of government regulators, such as FCC in the US and Ofcom in the UK etc., who proposed some important memoranda in succession to advocate a database-assisted spectrum access solution (FCC 2010, 2012; Ofcom 2015). In such a solution, TVWS devices obtain the available spectrum information through querying the geolocation database, instead of performing spectrum sensing. With the evolution of TVWS exploitation, it is increasingly recognized that exploiting geolocation databases coupled with spectral sensing allows a more efficient

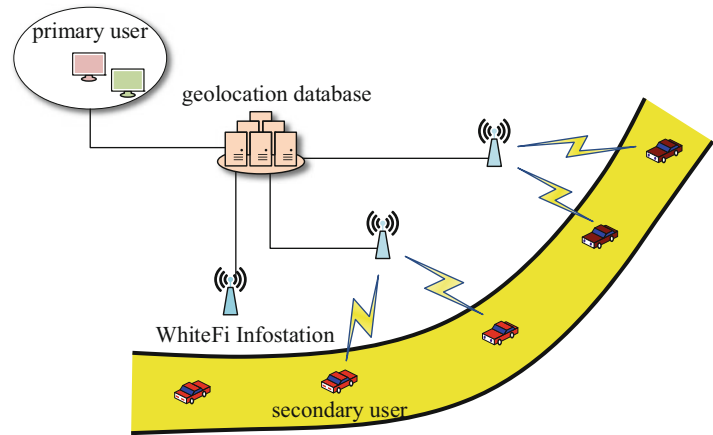
utilization of TVWS, especially when there exists geodatabase-unregistered Program Making and Special Event (PMSE) devices such as wireless microphones (Dionisio et al. 2014).

Foundations

TVWS is the unused or underutilized radio spectrum which used to be occupied by the licensed and protected users such as TV broadcast users and PMSE users. TVWS is also known as the VHF/UHF band locating at 470–790 MHz in Europe (Ofcom 2018; ETSI 2012) and noncontinuous 54–698 MHz in the United States (Electronic Code of Federal Regulations 2018). TVWS is now permitted to be shared by unlicensed users in their specific area without bringing interference to licensed users. Due to the abundant unlicensed spectrum resource at VHF/UHF bands and better signal penetration property for long-range wireless broadband access, TVWS exploitation has opened up a promising opportunity for massive wireless connectivity, e.g., vehicular ad hoc networks (VANET) (Flores et al. 2013). Although the database-assisted TVWS exploitation brings many benefits, the task of building up a secure and reliable database-assisted TVWS network system remains challenging (Murty et al. 2012). The TVWS network with WhiteFi Infostations (Zhou et al. 2016) to enable efficient vehicular content distribution for VANET was a feasible and efficient network system for reference, which overcame three challenges: (1) different vehicular media streaming applications of the TVWS secondary users have different quality of service (QoS) requirements, e.g., delay and throughput constraints; (2) the increasing number of the contending secondary users in the wide-range TVWS networks confront the severe medium access control (MAC) performance deterioration; and (3) vehicle mobility introduces significant dynamics to the long-range vehicular connection in TVWS. As depicted in Fig. 1, the geolocation database-assisted vehicular TVWS network was designed to be a location-aware, contention-free, and multi-polling vehicular

Database-Based Spectrum Access, Fig. 1

Geolocation database-assisted vehicular content distribution scenario



access scheme to satisfy the requirements of different vehicular media streaming applications. Like the general infrastructure-based TVWS networks, each WhiteFi Infostation first sends the required information such as its location and the transmission power to the geolocation database. The database then feeds back available vacant TV channels obtained from the primary users. Afterwards, each WhiteFi Infostation chooses the feasible channel to serve the secondary users, namely, the mobile vehicles within its transmission range.

Generally, there are two challenging issues in building up a database-assisted TVWS networks system: one is how to design the database and the other is how to price the spectrum. In the database design aspect, a robust and secure radio environment map database is required. The mathematical model applied to database design varies with application scenario. For example, the Markov decision process is suitable for the case where the spectrum lender has no knowledge of the spectrum borrower. Meanwhile, the Bayesian Stackelberg game may be a better choice when the spectrum lender has some knowledge about the spectrum borrower (Sodagari 2015). There are many rapidly changing application scenarios in TVWS networks which make the mathematical models complex and instable. As for the spectrum pricing aspect, the spectrum auction in a secondary market is a common pricing manner, where the secondary market means one in which the primary users could sell or lease idle spectrum

to many small secondary users, as opposed to the monolithic allocations in primary markets. There are many spectrum auction proposals in secondary market, such as SATYA (Kash et al. 2014), which take both the spectrum sharing and exclusive using into account, and the double-phase auction approach (Zhou et al. 2014) considering the diverse services priorities to improve the whole revenue. However, the auction-based manner brings high interaction overhead and thereby increases the communication delay. Therefore the more simple and efficient pricing strategies still need to be explored.

Key Applications

There are many potential applications for database-assisted TVWS access, such as the master-slave TVWS networks, offloading traffic to a TVWS network, TVWS serving as backhaul, rapid network deployment during emergencies, TVWS used for local TV broadcaster, etc. (RFC6953 2018).

Cross-References

- ▶ [Dynamic Spectrum Access Through Cognitive Radio Networking](#)
- ▶ [Spectrum Resource Allocation in Cognitive Radio Networks](#)
- ▶ [Spectrum Sensing in Cognitive Radio Networks](#)

References

- Cordeiro C, Challapali K, Birru D, Shankar S (2006) IEEE 802.22: an introduction to the first wireless standard based on cognitive radios. *J Commun* 1(1):38–47
- Dionisio R, Ribeiro J, Marques P, Rodriguez J (2014) Combination of a geolocation database access with infrastructure sensing in TV bands. *EURASIP J Wirel Commun Netw* 2014(1):210
- Electronic Code of Federal Regulations. Title 47, part 15, subpart H, television band devices. <http://www.ecfr.gov/>. Accessed May 2018 from GPO
- ETSI (2012) EN 301 598 white space devices (WSD), wireless access systems operating in the 470 MHz to 790 MHz frequency band. Oct 2012
- FCC (2010) Second memorandum opinion and order. FCC, Washington, DC
- FCC (2012) Third memorandum opinion and order. FCC, Washington, DC
- FCC. FCC frees up vacant TV airwaves for “Super Wi-Fi” technologies and other technologies. <https://www.fcc.gov/general/statements-meredith-attwell-baker>. Accessed May 2018
- Flores A, Guerra R, Knightly E, Ecclesine P, Pandey S (2013) IEEE 802.11af: a standard for TV white space spectrum sharing. *IEEE Commun Mag* 51(10):92–100
- Kash I, Murty R, Parkes DC (2014) Enabling spectrum sharing in secondary market auctions. *IEEE Trans Mob Comput* 13(3):556–568
- Murty R, Chandra R, Moscibroda T, Bahl P (2012) Senseless: a database-driven white spaces network. *IEEE Trans Mob Comput* 11(2):189–203
- Ofcom (2015) Implementing TV white spaces. Ofcom, London. Technical report 08-260
- Ofcom. Regulatory requirements for white space devices in the UHF TV band. <http://www.cept.org/Documents/se-43/6161/>. Accessed May 2018
- RFC6953. Protocol to access white-space (PAWS) databases: use cases and requirements. <http://www.rfc-editor.org/info/rfc6953>. Accessed May 2018
- Sodagari S (2015) A secure radio environment map database to share spectrum. *IEEE J Sel Top Sign Proces* 9(7):1298–1305
- Zhou H, Liu B, Hou F, Zhang N, Gui L, Chen J, Shen X (2014) Database-assisted dynamic spectrum access with QoS guarantees: a double-phase auction approach. In: *Proceedings of the 3th IEEE/CIC international conference on communications in China*, pp 66–77
- Zhou H, Cheng N, Lu N, Gui L, Zhang D, Yu Q, Bai F, Shen X (2016) WhiteFi infostation: engineering vehicular media streaming with geolocation database. *IEEE J Sel Areas Commun* 34(8):2260–2274

Data-Centric Wireless Sensor Networks

► Information-Centric Wireless Sensor Networks

Data-Driven Digital City

► Data-Driven Smart City

Data-Driven Edge Computing

Siguang Chen

Nanjing University of Posts and Telecommunications, Nanjing, China

Synonyms

Edge computing; Fog computing; Local cloud; Mobile edge computing

Definition

Data-driven edge computing (also similar with local cloud and fog computing) (Lopez et al. 2015; Satyanarayanan et al. 2009; Bonomi et al. 2012; Milan et al. 2014) is a model to complement cloud computing systems by performing data processing at the edge of the network, near the source of the data. As the era of big data has arrived, billions of connected devices generating petabytes of data will demand nearby computing resources to provide real-time (or low latency) data services, edge computing decentralizes the concentration of computing resources and brings the computing closer to the devices requesting that computing power, which can improve the quality of service (QoS) and user experience significantly.

Historical Background

With the rapid and continuous deployment of enormous number of data sensing devices (such as sensors, smart meters, smart glasses, smart phones, smart vehicles, etc.), the data scale of network is showing a growth spurt. Cisco predicts

that 50 billion devices will connect to the Internet by 2020 (Evans 2011), this number will reach 500 billion by 2025 (Camhi 2015). By 2019, the data produced by people, machines, and things will reach 500 zettabytes (Cisco 2014). The large-scale data often accompanied by the great burden of data processing, collection, and storage. The traditional architectures of wireless sensor network (WSN), ad hoc, and Internet of Things (IoTs) cannot satisfy the requirement of efficient data processing. How to relieve the processing pressure of big data and improve the QoS becomes a great challenge.

Edge computing is usually deployed at local region and used to compensate drawbacks of cloud computing, which can utilize the computation capabilities of edge (local) devices effectively. The preprocessing operation of edge device can alleviate the data processing burden of cloud node and reduce the communication overhead of network significantly (Chen et al. 2018). Overall, the employing of fog computing can reduce computation latency, bandwidth, and energy consumptions significantly. For the sake of effectively using the edge device's computation and storage resources, improving the QoS and user experience, numerous edge computing-based schemes are developed to achieve such goal.

Latency is an important performance metric for user experience. The work of Rimal et al. (2016) investigated performance gains of centralized cloud and mobile edge computing (MEC) enabled integrated fiber-wireless (FiWi) access networks, which shows the offload packet delay can be reduced without affecting network performance of broadband access traffic. Varghese et al. (2017) highlighted the feasibility and the benefits in improving the QoS and experience by using fog computing. The experimental results demonstrate that the average response time for a user is improved by 20% when using the edge of the network in comparison to using a cloud-only model. Meanwhile, the volume of traffic between the edge and the cloud server is reduced by over 90% for the use case.

Bandwidth consumption is an important performance metric for QoS. From latency

perspective, low bandwidth consumption can reduce the network congestion and then reduce transmission time, especially for large data (e.g., video stream, etc.). Edge computing urges the video stream which don't need to route to the central cloud, it avoids a great amount of network bandwidth consumption due to the nature of video stream, and also avoids the problem of network congestion caused by the video stream uploading. The work of Mehta et al. (2016) studied the benefit of introducing additional data centers closer to the network edge for the optimal application placement. The study shows that the cost savings for such operation can go up to 67% for the bandwidth-hungry applications. The similar research results indicate that backhaul savings can go up to 22% by exploiting the proactive caching scheme (Bastug et al. 2014).

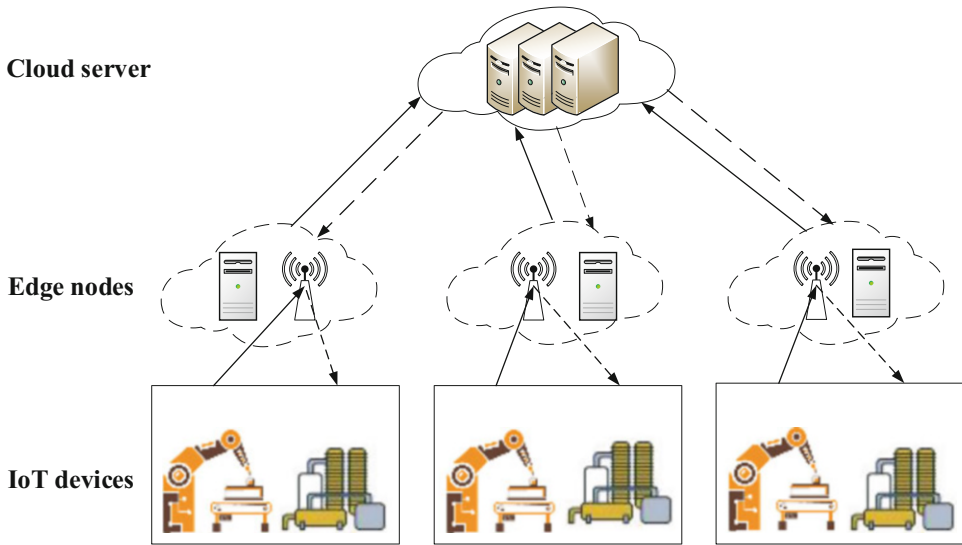
Energy consumption is an important performance metric for network efficiency. Battery is the most precious resource for devices at the edge of the network; the research work shows that the energy consumption in the device can be reduced by up to 42% through offloading to edge servers compared to cloud offloading (Gao et al. 2015). Zhang et al. (2016) developed an energy-efficient computation offloading (EECO) mechanisms for MEC in 5G heterogeneous networks. The proposed mechanism jointly optimizes offloading and radio resource allocation to obtain the minimal energy consumption under the latency constraints.

For detailed information on the background and state of the art of edge computing, please read the review works of Wang et al. (2017) and Ni et al. (2018).

Foundations

The following figure shows a typical data-driven edge-cloud computing model for Internet of Things (IoT) systems.

As shown in Fig. 1, a hierarchical IoT computing system consisting of one remote cloud server, several edge nodes, and a number of IoT devices is constructed. Each IoT device with a volume of data is connected to the nearby edge node



Data-Driven Edge Computing, Fig. 1 Data-driven edge-cloud computing model

through the wireless access point. The edge node receives the data processing requests uploaded by the nearby devices and will return the results to the terminal devices after these requests are processed. The edge node can provide timely data processing services for time-sensitive applications. Meanwhile, when the edge node is overloaded, it can offload the partial computation tasks to the cloud server for improving the user experience and extending the finite computation and storage capabilities of edge node. This hybrid edge-cloud computing paradigm can make full use of the advantages of both edge and cloud computing modes and meet the data processing requirements of big data scenario.

Key Applications

The advantages of edge computing urge that can be applied in various fields, such as dynamic content delivery, video analytics, IoTs, vehicular networks, big data analytics, etc.

1. Dynamic content delivery

In the edge computing paradigm, not only data should be cached at the edge server but also it can perform operations on the data. Caching and processing at the edge server can

provide dynamic content delivery based on the network information (such as network load, network status, etc.) and user's context-aware information (Zhu et al. 2013). One typical application is online shopping services for the frequent changes of shopping cart. Edge-based dynamic content delivery improves the quality of experience (QoE) for approaching to end users effectively.

2. Video analytics

With the continuous increasing of video monitoring devices and network cameras, the video analytics become an emerging technology. The use of edge computing paradigm allows the video analysis at any edge devices and gets the result much faster compared with traditional cloud computing. The user requests of data analysis perform at local devices which avoid numerous redundant video streams transport through the core network to cloud server. Traditional cloud computing is no longer suitable for video analytics application due to the high latency of long distance transmission and privacy concerns.

3. Internet of Things

The application fields of edge computing paradigm in IoTs mainly include smart health-

care, smart grid, smart home, smart transportation, etc. In healthcare applications, real-time processing and emergent event response are very important for stroke patients and sudden cardiac death. Edge computing paradigm can definitely shorten the time of first aid and improve the success rate in rescuing patients (Ni et al. 2018). Similarly, edge computing can provide efficient energy consumption collection scheme for smart grid, offer low latency, localized, and plug and play services for smart home (Wang et al. 2017), and improve traffic jam and safety for vehicles and pedestrian.

4. Vehicular networks

Nowadays, a vehicle can communicate with other vehicles (V2V), surrounding infrastructure (V2I), the Internet (V2N), roadside pedestrian (V2P), and a back-end cloud server (Liu et al. 2016); it means that the vehicle can communicate with everything (V2X). V2X can achieve a series of new applications, such as automated overtake, autonomous driving, cooperative collision avoidance, bird's eye view, traffic prediction, and smart parking (Wan et al. 2016; Han et al. 2016; Liu et al. 2017). MEC can reduce the transmission latency and cost of the computation offloading for the close proximity to the mobile vehicular terminals.

5. Big data analytics

With the rapid increasing of data scale, big data analytics becomes a huge challenge for traditional cloud-based network, since the data collection from the terminal devices are transmitted to the remote cloud server for big data analytics. Edge computing can manage and use the local computing and storage resources, and is a promising choice to handle big data for which can reduce the bandwidth consumption and improve response time.

Cross-References

- [Data-Driven QoE Analysis](#)
- [Fog Computing for Intelligent Mobile and IoT Networks](#)

- [Mobile Edge Caching in HetNets](#)
- [Wireless Edge Caching](#)

References

- Bastug E, Bennis M, Debbah M (2014) Living on the edge: the role of proactive caching in 5G wireless networks. *IEEE Commun Mag* 52(8):82–89
- Bonomi F, Milito R, Zhu J et al (2012) Fog computing and its role in the internet of things. In: *Proceedings of the 1st ACM Mobile Cloud Computing workshop*, pp 13–15
- Camhi J (2015) Former Cisco CEO John Chambers predicts 500 billion connected devices by 2025. *Business Insider*
- Chen S, Du L, Wang K et al (2018) Fog computing based optimized compressive data collection for big sensory data. In: *Proceedings of the IEEE ICC*, pp 1–6
- Cisco (2014) Cisco global cloud index: forecast and methodology, 2014–2019. Cisco White Paper
- Evans D (2011) The internet of things: how the next evolution of the internet is changing everything. Cisco White Paper
- Gao Y, Hu W, Ha K et al (2015) Are cloudlets necessary? Technical Report CMU-CS-15-139. School of Computer Science, Carnegie Mellon University, Pittsburgh
- Han G, Jiang J, Guizani M et al (2016) Green routing protocols for multimedia sensor networks. *IEEE Wirel Commun* 23(6):140–146
- Liu J, Wan J, Wang Q et al (2016) A survey on position-based routing for vehicular ad hoc networks. *Telecommun Syst* 62(1):15–30
- Liu J, Wan J, Zeng B et al (2017) A scalable and quick-response software defined vehicular network assisted by mobile edge computing. *IEEE Commun Mag* 55(7):94–100
- Lopez PG, Montresor A, Epema D et al (2015) Edge-centric computing: vision and challenges. *SIGCOMM Comput Commun Rev* 45(5):37–42
- Mehta A, Tärneberg W, Klein C et al (2016) How beneficial are intermediate layer data centers in mobile edge networks?. In: *Proceedings of the IEEE 1st International Workshops on Foundations and Applications of Self-Systems (FAS-W)*, pp 222–229
- Milan P, Jerome J, Valerie Y et al (2014) Mobile-edge computing-introductory technical white paper. White Paper, ETSI, Sophia Antipolis
- Ni J, Zhang K, Lin X et al (2018) Securing fog computing for internet of things applications: challenges and solutions. *IEEE Commun Surv Tutor* 20(1):601–628
- Rimal BP, Van DP, and Maier M (2016) Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks. In: *Proceedings of the IEEE conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp 991–996
- Satyanarayanan M, Bahl P, Caceres R et al (2009) The case for VM-based cloudlets in mobile computing. *IEEE Pervasive Comput* 8(4):14–23

- Varghese B, Wang N, Nikolopoulos DS et al (2017) Feasibility of fog computing. arXiv preprint arXiv:1701.05451
- Wan J, Liu J, Shao Z et al (2016) Mobile crowd sensing for traffic prediction in internet of vehicles. *Sensors* 16(1):1–15
- Wang S, Zhang X, Zhang Y et al (2017) A survey on mobile edge networks: convergence of computing, caching and communications. *IEEE Access* 5:6757–6779
- Zhang K, Mao Y, Leng S et al (2016) Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks. *IEEE Access* 4:5896–5907
- Zhu J, Chan DS, Prabhu MS et al (2013) Improving web sites performance using edge servers in fog computing architecture. In: *Proceedings of the IEEE 7th International symposium on Service-Oriented System Engineering (SOSE)*, pp 320–323

Data-Driven Intelligent City

► Data-Driven Smart City

Data-Driven Mobile Social Networks

Fuliang Li
Northeastern University, Shenyang, China

Synonyms

Mobile big data; Mobile social network

Definitions

Mobile social networks (MSNs) combine the social networks with the mobile communication networks. In MSN, users with similar interests communicate with one another through their mobile devices. Data-driven MSNs take full advantages of mobile big data analysis to improve the design and optimize the performance of MSNs.

Historical Background

Online social networks (OSNs, e.g., Ins, Facebook, and Twitter) provide a platform for the people with the similar interests and background and to communicate with one another (Schneider et al. 2009). Most OSNs are the web-based applications. Some studies have confirmed that there are positive psychological outcomes from OSNs compared with the traditional offline social networks (Oh et al. 2014). However, with the development of wireless access technology, as well as the wide popularity of mobile devices, mobile users increase sharply. As a result, the traditional PC end users are replaced by the mobile end users gradually in OSNs. To provide better services for mobile users, mobile social networks (MSNs) have emerged, which leverage the characteristics of social networks in combination with mobile communication networks (Hu et al. 2015). On one hand, with the aid of the social measurements in social networks, the routing mechanism in mobile communication networks can be optimized, and thus the quality of experience is improved as well. On the other hand, social networks can exploit the ubiquitous accessibility of mobile communication networks to realize real-time interaction even without connections to the Internet. In our daily life, many MSN applications have been developed based on the location and sensing, such as Geo-Life (Zheng et al. 2009), Foursquare (Foursquare 2018), and Micro-blog (Gaonkar et al. 2008), etc.

Unlike traditional OSNs, MSNs fully utilizes the characteristics of mobile devices, such as portability and multiple sensing capabilities (Mao et al. 2017). On one hand, owing to the portability, mobile users can access any social platforms anywhere at any time. On the other hand, mobile devices integrate a variety of sensors, such as light sensor, accelerator sensor, gyroscope, temperature sensor, blood pressure sensor and heart rate sensor, etc. The multiple sensing capabilities strengthen the relations between MSNs and mobile users. There are three typical architectures of MSNs listed as follows.

- *Centralized MSN.* It follows the C/S mode, and all data are stored in the centralized servers. Communications among the mobile devices are based on the Internet through wireless access technologies, such as 4G and Wi-Fi, etc.
- *Distributed MSN.* In this architecture, none of centralized servers are utilized. Mobile users directly communicate with one another based on the device-to-device (D2D) communication technologies, such as Bluetooth (Bluetooth 2018) and Wi-Fi Direct (WFD 2016), etc.
- *Hybrid MSN.* This is a hybrid architecture of the above two cases. It supports the infrastructure-based communications (i.e., centralized MSN), as well as the infrastructure-less communications (i.e., distributed MSN). The hybrid MSN is a flexible solution, and it depends on the practical MSN-based applications

The social attribute is a typical characteristic of MSNs. To measure the social attributes, social behavior analysis is introduced (Hu et al. 2015). Social behavior can be derived from the social relationship (Li and Chen 2009) and the social influence (Peng et al. 2017). Additionally, the capability of context awareness of users is a major advantage of MSNs (Mao et al. 2017). With the help of the embedded sensors in mobile devices, MSNs can capture the contextual information, for example, the location information without user intervention. Based on the contextual information, the functions of MSN applications can be extended and diversified (Araniti et al. 2017; Yu et al. 2017). However, how to protect the user privacy in MSNs becomes a challenging issue. Li et al. (2018) proposed a novel attack that utilizes the information of users' location to infer the demographic attributes. And then they presented a privacy control approach called SmartMask to defense the attack. Wang et al. (2018) proposed a two-hop feedback-based dynamic trust framework to recognize the selfish and malicious nodes in opportunistic MSNs, which could defense the conspiracy attacks

effectively. To ensure the security and efficiency of content caching in fog computing enabled MSNs, Su et al. (2018) proposed a secure content caching schema following the idea of disaster backup.

In MSNs, researchers and developers require to acquire lots of data from mobile devices. As mobile traffic will increase sevenfold between 2016 and 2021 (Cis 2017), mobile data already have the typical characteristics of big data, such as large volume and economic value (Hu et al. 2014). By analyzing the mobile data, some valuable information could be obtained to improve the design and optimize the performance of MSNs. Whereas how to reasonably store, manage, deliver, and control these mobile big data has become a challenging issue (Su et al. 2016), which promotes the evolution of data-driven MSNs.

In fact, the data-driven is an idea, which can be implemented by some specific approaches, such as data mining and machine learning, etc. Du et al. (2011) used the standard Markov model and the semi-Markov model to predict the user mobility. In opportunistic networks, the connectivity of nodes is discontinuous, and the mobility of nodes is driven by human sociality. To improve the communication success rate, a sociality-driven communication method was introduced (Societevole et al. 2015), which could reflect the social behaviors of nodes and discover available relays for message forwarding. It is effective to adopt the data-driven thought to characterize the user mobility in a wide range of mobile computing applications. Yuan et al. (2016) proposed a model based on data analysis to identify unlicensed taxis. The model firstly screened out a coarse-grained suspected unlicensed taxi candidate list. Then, the results were used to get a fine-grained list of suspected unlicensed taxis based on machine learning.

There are many studies on optimizing MSNs based on the data-driven thought. Zhang et al. (2018) conducted traffic prediction based on the data gathered from WeChat Moments. They also discussed how to develop new MSN applications with the help of the mobile big data.

Wu et al. (2017) collected some user information from 17000 mobile devices. Results revealed that the social relationship has a great impact on tweet click behaviors. In addition, they also presented a cluster-based Latent Bias Model for data-driven learning-based prefetching prediction in MSNs. Data-driven modeling should be conducted based on big data to ensure the correctness and the effectiveness. Wang et al. (2017) pointed out that limited data analysis caused by small scale of dataset or single-dimensional feature may restrict the applications in practice severely. And they analyzed the behaviors of 30 million users from the platform of Xender for device-to-device content sharing in MSNs. Song et al. (2015) proposed a model based on Bayes' rule to predict the user locations in social networks by social influence from the perspectives of temporal and spatial, respectively. They found that, in location-based social networks, the social influence plays an important role in modeling user mobility through analyzing the check-in movement behaviors of users. Shi et al. (2016) investigated the impacts of space and place on social networks using mobile phone data. They confirmed the distance decay effects in social networks and found that individuals interaction is a more important factor than spatial proximity.

Given above, the data-driven thought can take full advantages of mobile big data analysis to promote the development of MSNs. Through data-driven modeling, the existing designs of MSNs such as routing and data sharing could be improved and optimized. Therefore, data-driven MSN has become a promising field in the current research on MSNs.

Key Applications

Data-driven MSNs take full advantages of mobile big data analysis to improve the design and optimize the performance of MSNs, which could act the basis for offloading computing and edge computing.

Cross-References

- [Data-Driven Edge Computing](#)
- [Data-Driven Resource Allocation](#)
- [Data-Driven Smart City](#)

References

- Araniti G, Orsino A, Militano L, Wang L, Iera A (2017) Context-aware information diffusion for alerting messages in 5g mobile social networks. *IEEE Internet Things J* 4(2):427–436
- Bluetooth (2018) Bluetooth. <https://www.bluetooth.com/>
- Cis (2017) Cisco visual networking index: global mobile data traffic forecast update, 2016–2021 White Paper. Cisco
- Du Y, Fan J, Chen J (2011) Experimental analysis of user mobility pattern in mobile social networks. In: *Wireless communications and networking conference (WCNC)*, 2011 IEEE. IEEE, pp 1086–1090
- Foursquare (2018) Foursquare. <https://foursquare.com/>
- Gaonkar S, Li J, Choudhury RR, Cox L, Schmidt A (2008) Micro-blog: sharing and querying content through mobile phones and social participation. In: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, pp 174–186
- Hu H, Wen Y, Chua TS, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. *IEEE Access* 2:652–687
- Hu X, Chu TH, Leung VC, Ngai ECH, Kruchten P, Chan HC (2015) A survey on mobile social networks: applications, platforms, system architectures, and future research directions. *IEEE Commun Surv Tutor* 17(3):1557–1581
- Li H, Zhu H, Du S, Liang X, Shen XS (2018) Privacy leakage of location sharing in mobile social networks: attacks and defense. *IEEE Trans Dependable Secure Comput* 15(4):646–660
- Li N, Chen G (2009) Multi-layered friendship modeling for location-based mobile social networks. In: *Mobile and ubiquitous systems: networking & services, MobiQuitous, 2009. MobiQuitous'09. 6th annual international*. IEEE, pp 1–10
- Mao Z, Jiang Y, Min G, Leng S, Jin X, Yang K (2017) Mobile social networks: design requirements, architecture, and state-of-the-art technology. *Comput Commun* 100:1–19
- Oh HJ, Ozkaya E, LaRose R (2014) How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Comput Human Behav* 30: 69–78

- Peng S, Yang A, Cao L, Yu S, Xie D (2017) Social influence modeling using information theory in mobile social networks. *Inf Sci* 379:146–159
- Schneider F, Feldmann A, Krishnamurthy B, Willinger W (2009) Understanding online social network usage from a network perspective. In: *Proceedings of the 9th ACM SIGCOMM conference on internet measurement*. ACM, pp 35–48
- Shi L, Wu L, Chi G, Liu Y (2016) Geographical impacts on social networks from perspectives of space and place: an empirical study using mobile phone data. *J Geogr Syst* 18(4):359–376
- Socievole A, Caputo A, Marano S (2015) Multi-layer sociality in opportunistic networks: an extensive analysis of online and offline node social behaviors. In: *2015 international symposium on performance evaluation of computer and telecommunication systems (SPECTS)*. IEEE, pp 1–8
- Song Y, Hu Z, Leng X, Tian H, Yang K, Ke X (2015) Friendship influence on mobile behavior of location based social network users. *J Commun Netw* 17(2):126–132
- Su Z, Xu Q, Qi Q (2016) Big data in mobile social networks: a qoe-oriented framework. *IEEE Netw* 30(1):52–57
- Su Z, Xu Q, Luo J, Pu H, Peng Y, Lu R (2018) A secure content caching scheme for disaster backup in fog computing enabled mobile social networks. *IEEE Trans Ind Inform* 14(10):4579–4589
- Wang EK, Li Y, Ye Y, Yiu SM, Hui LC (2018) A dynamic trust framework for opportunistic mobile social networks. *IEEE Trans Netw Serv Manag* 15(1):319–329
- Wang H, Wang S, Zhang Y, Wang X, Li K, Jiang T (2017) Measurement and analytics on social groups of device-to-device sharing in mobile social networks. In: *2017 IEEE international conference on communications (ICC)*. IEEE, pp 1–6
- WFD (2016) Wi-fi peer-to-peer (p2p) technical specification v1.7. Technical report, WiFi Alliance
- Wu C, Chen X, Zhu W, Zhang Y (2017) Socially-driven learning-based prefetching in mobile online social networks. *IEEE/ACM Trans Netw (TON)* 25(4):2320–2333
- Yu Z, Zhang D, Wang Z, Guo B, Roussaki I, Doolin K, Claffey E (2017) Toward context-aware mobile social networks. *IEEE Commun Mag* 55(10):168–175
- Yuan W, Deng P, Taleb T, Wan J, Bi C (2016) An unlicensed taxi identification model based on big data analysis. *IEEE Trans Intell Transp Syst* 17(6):1703–1713
- Zhang Y, Li Z, Gao C, Bian K, Song L, Dong S, Li X (2018) Mobile social big data: Wechat moments dataset, network applications, and opportunities. *IEEE Netw* 32(3):146–153
- Zheng Y, Chen Y, Xie X, Ma WY (2009) Geolife2.0: a location-based social networking service. In: *Tenth international conference on mobile data management: systems, services and middleware, 2009 (MDM'09)*. IEEE, pp 357–358

Data-Driven Network Control

An Xie and Xiaoliang Wang

Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu Province, China

Background

Traditionally, network control is usually optimized for better experience of applications with the goal of avoiding congestion or reducing flow completion time in data center. To this end, enormous amount of researches have focused on adjusting parameters of certain transmission protocols, such as TCP, DCTCP, D2TCP, etc. These works can be summarized as “optimizing applications by adjusting network control.”

With the emerging of big data, tremendous amount of data is generated in end devices or by data applications in data center. As a result, how to leverage the benefit of this huge amount of data to better control the network has attracted research communities’ attention. The new paradigm, referred to as “Data-Driven network Control,” aims at improving network performance by leveraging user/applications-generated data. Compared to the traditional “optimizing applications by adjusting network control,” data-driven network control can be identified in the opposite direction, i.e., “optimization network control by data of applications.”

In the following, we will first introduce some basic concepts and mechanism of data-driven network control. Then we will compare the traditional model-based control with data-driven control to highlight unique properties of data-driven control. At last, some preliminary applications are given to show potential benefits of data-driven network control.

Data-Driven Network Control

The main idea of data-driven network control is to exploit large volumes of network data to

optimize network performance. Most network control mechanisms nowadays use only local data of a single flow. In contrast, data-driven network control uses data from multiple flows or even the network to make better control decisions.

The Opportunity of Using Big Data for Network Control

Tremendous network devices and users are generating more data than ever before. For example, according to the report of IBM, 2.5 quintillion bytes of data are created everyday (Wu et al. 2014). Big data is usually characterized by five V's: (1) volume, i.e., the size of data is huge; (2) velocity, i.e., the data production rate is high; (3) variety, i.e., the type of data is various; (4) veracity, i.e., the quality of data is high; (5) value, i.e., the data is in general useful. It is notable that the produced data sets are so large and complex that traditional methods are inadequate to accurately model and analyze them. Therefore, people may refer to new technologies such as data mining and machine learning.

Definition of Data-Driven Network Control

We borrow the definition from (Jiang et al. 2017) as follows. Data-driven network is a paradigm for designing the control plane of network protocols. Data-driven network contains two components: the client-side instrumentation and the controller. The client-side instrumentation is responsible for measuring client-perceived quality and applying decisions made by the controller while the controller is responsible for aggregating measurements and making control decisions.

Mechanism of Data-Driven Network Control

We introduce the mechanism of data-driven network control on the three-layer architecture from (Yao et al. 2016) which contains infrastructure layer, control layer, and application layer.

- *The infrastructure layer.* This is where the client-side instrumentation resides. It may contain various devices, such as switches,

access points, and smart phones. Devices in the infrastructure layer usually implement protocols through which the control layer can fetch statistics or logs. For example, the OpenFlow protocol in SDN allows controller to query flow statistics. Devices in the infrastructure layer also accept orders from the control layer to alter network configuration.

- *The control layer.* The controller is in this layer. This layer resides between the infrastructure layer and the application layer. The main responsibility of the control layer is to analyze data collected from the infrastructure layer. As stated above, in general, both the number of devices and the amount of data are huge. Therefore, data-driven approaches like machine learning and data mining are advantageous over the model based approaches.
- *The application layer* contains user applications and it resides on top of the control layer. Applications obtain models generated by the control layer. In conjunction with the network management protocols such as SNMP or SDN on the control layer, applications can maximize their performance by dynamically changing network parameters.

Model-Based Control Versus Data-Driven Control

To design a control system from data, there are mainly two approaches, i.e., the model-based approach and the data-driven approach. The model-based approach first derives a model from the data after which a controller is computed based on the model. In contrast, the data-driven approach directly derives the controller from the data.

Model-Based Control (MBC)

The model-based control requires that the parameters be calibrated using collected data that minimizing the model reference criterion or satisfying some user-defined requirements. The derived model either reflects or approximates the true

system. According to (Murray 2010), the model-based design has the following advantages:

- Model-based design provides a common design environment, which facilitates general communication, data analysis, and system verification between various groups.
- Engineers can locate and correct errors early in system design, when the time and financial impact of system modification are minimized.
- Design reuse, for upgrades and for derivative systems with expanded capabilities, is facilitated.

The main problems of this approach are that the performance strongly depends on the modeling technique and is limited by the modeling errors.

Data-Driven Control (DDC)

Sometimes it not possible and feasible to get the accurate model of objects. Reasons are twofold. First, it may be difficult to get the accurate model of every object. For example, in Internet of Things (IoT), there are a large number of devices. Second, it may be too difficult to model even one object. This is because that many industry processes, aerospace systems, transportation systems, power grid systems, etc., are becoming more and more complex. Fortunately, in many cases, collecting data is relatively easy thanks to the development of information technology and the decrease in the price of storage media. Moreover, it may be the only way we know about objects in IoT.

Data-driven control aims at making full use of those data, either collected online or offline. It directly derives the control system from the data instead of first modeling the plant and then designing the controller using the plant model. At the same time, properties such as stability, convergence, and robustness can be guaranteed under certain assumptions. The main advantages of data-driven control are summarized as follows (Hou and Wang 2013):

- The controller in DDC approaches does not explicitly include any parts or the whole of the plant model. For this reason, it has overcome the dependence on the plant model for the design of control systems.
- The stability and convergence conclusions of DDC approaches do not depend upon the accuracy of the model, excepting the DDC control methods implicitly using the system dynamics and structure information, such as direct adaptive control, subspace predictive control, which is the main stumbling block for the applications of the MBC theory.
- The most outstanding point of DDC approaches is that the twinborn problem of un-modeled dynamics and robustness in traditional MBC theory do not exist under DDC framework.

In the following section, we review some illustrative applications based on data-driven network control paradigm.

Applications

Better CDN Selection

The observed video quality delivered by individual CDNs can vary substantially across clients and time. As the video player has only a few seconds to do the buffering, decisions need to be made quickly to let the buffer not drain out. C3 (Ganjam et al. 2015) is one data-driven approach of dynamic CDN selection by exploiting quality measurements of other video sessions. C3 is composed of two layers, i.e., a global controller and a client side sensing/actuation layers. The controller uses real-time measurements of sensing/actuation layer. These measurements are used by the controller to create a model of expected performance and in turn, this model is used by the controller to decide suitable parameters such as CDN and bitrate for clients. The decisions are executed on the sensing/actuation layers. C3 has been operational for 8 years, and has improved the video quality significantly.

Adaptive Bitrate Selection

Adaptive bitrate selection is critical to ensure good Quality of Experience (QoE) for Internet video. To meet the increasing stringent QoE demand of users, video players need intelligent bitrate selection and adaptation algorithms. To this end, accurate throughput prediction plays a vital role in designing the two algorithms. CS2P (Sun et al. 2016) analyzes throughput characteristics in a dataset with 20M+ sessions from one of the biggest video providers in China, iQIYI. CS2P reveals that sessions sharing similar key features such as ISP and region have similar initial throughput values and dynamic patterns. Besides, a natural stateful behavior exists in throughput variability within a given session.

Based on the findings, CS2P develops an offline clustering based prediction model using observed throughput from past video sessions. The model clusters sessions that are likely to exhibit similar throughput patterns. CS2P determines the initial throughput for each cluster by the median throughput. Moreover, CS2P learns a Hidden-Markov-Model (HMM) for each cluster to model the stateful evolution of intrasession throughput. The two models are then plugged into the video player and determine the initial bitrate and midstream bitrate adaption online. CS2P achieves 3.2% improvement on overall QoE and 10.9% higher average bitrate over state-of-art Model Predictive Control (MPC) approach.

References

- Ganjam A, Siddiqui F, Zhan J, Liu X, Stoica I, Jiang J, Sekar V, Zhang H (2015) C3: Internet-scale control plane for video quality optimization. In: 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2015), Santa Clara, CA, pp. 131–144.
- Hou Z-S, Wang Z (2013) From model-based control to data-driven control: survey, classification and perspective. *Inf Sci* 235:3–35
- Jiang J, Sekar V, Stoica I, Zhang H (2017) Unleashing the potential of data-driven networking. In: International Conference on Communication Systems and Networks. Springer, Bangalore, India, pp 110–126
- Murray CJ (2010) Automakers opting for model-based design. *Des News* 5(11)
- Sun Y, Yin X, Jiang J, Sekar V, Lin F, Wang N, Liu T, Sinopoli B (2016) Cs2p: improving video bitrate selection and adaptation with data-driven throughput prediction. In: Proceedings of the 2016 ACM SIGCOMM conference, Florianópolis, Brazil, ACM, pp 272–285
- Wu X, Zhu X, Gong-Qing W, Ding W (2014) Data mining with big data. *IEEE Trans Knowl Data Eng* 26(1):97–107
- Yao H, Qiu C, Fang C, Chen X, Yu FR (2016) A novel framework of data-driven networking. *IEEE Access* 4:9066–9072

Data-Driven Pattern Prediction

Zhuzhong Qian

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Synonyms

[Data Mining](#); [Data Analytics](#)

Definition

Data-driven pattern prediction is the process of using large volume of related historical data to extract the inner patterns of some observed objects on their behaviors or some phenomena, and then using extracted patterns to make accurate prediction rapidly.

Key Application

- Guiding keyword bidding based on the patterns of user clicks from web logs.
- Guiding traffic evacuation based on the patterns of rush hours in a city.
- Guiding resource management based on the patterns of resource demands in a cluster.

Background of Data-Driven Pattern Prediction

With the rapid growth of the Internet, there is a huge amount of data accumulated in human's activities. For example, within the Internet minute in 2013 (Lena long 2013), there are tens of thousands of application downloads, millions of search queries, and tens of millions of photo views. As a result, hundreds of thousands of gigabytes are transferred globally through IP packets over the network within an Internet Minute. Apart from those network activities, millions of cameras have already deployed around the world (Zhang et al. 2017), which would also generate large volume of data for surveillance and business intelligence.

Such large volume of data, also well known as big data (Big Data 2016), can be used for predictive analytics, user behavior analytics, or certain other advanced data analytics methods by extracting valuable information from data. The new paradigm, referred to as "data-driven analysis," aims at digging valuable information from the big data to help improve the performance of the network, enhance the surveillance, spur the business intelligence, etc. Note that there are many inner patterns existing within big data. For example, as shown in Fig. 1, the population of Manhattan changes with patterns. The data is collected from the Metropolitan Transportation Authority's turnstile database (Turnstile Data of Metropolitan Transportation Authority 2018) (nearly 10GB data) and Steven Romalewski's MTA subway data (Mta Subway Data 2018). Although the population changes hour by hour, there are periodical waves between two different days. This is because the tidal routine of people occurs day by day. Therefore, if we can extract the patterns and make prediction based on the changes of population from the big data of subway, we have more opportunities to improve the public transportation or solve other related problems.

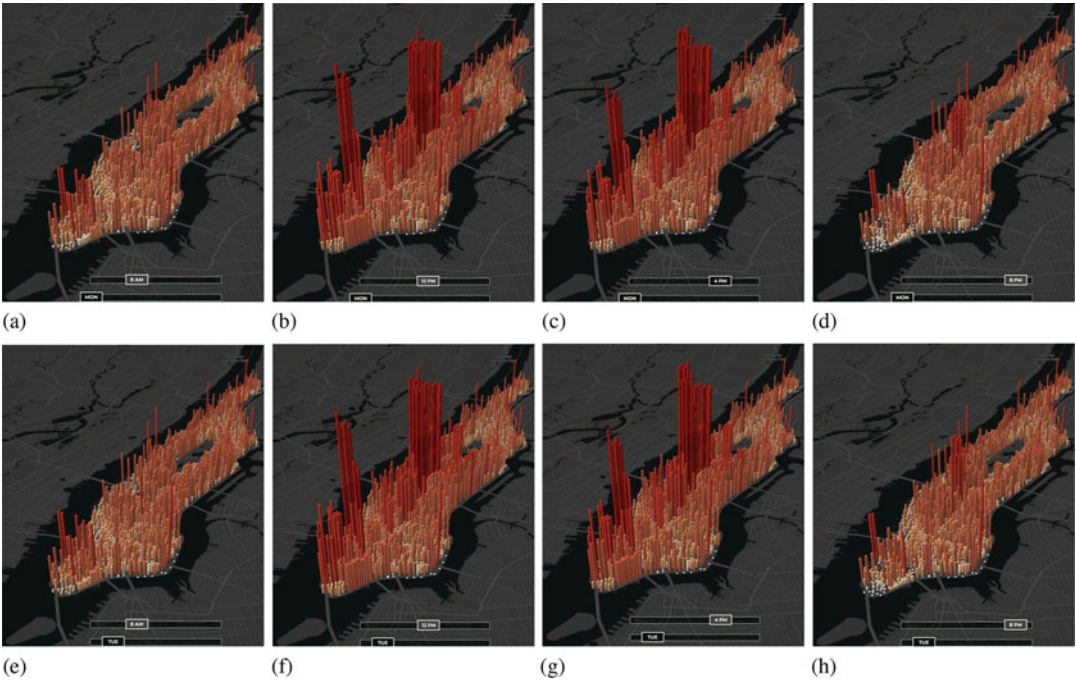
Apart from the public transportation, we can extract the valuable patterns from data and make prediction for wide range of scenarios. For exam-

ple, we can use the logs from websites/sessions to capture the browsing patterns of users, guiding better user experience. We can also use the resource demands from various applications during their executions to guide the efficient resource management within clouds, edges, and so on.

Data-Driven Pattern Prediction

Here, the process of extracting valuable information from data is called "data-driven." Due to the existence and value of inner patterns inside big data, extracting patterns and making prediction from big data is urgent for data analysts. But, how to extract the patterns from big data is a big challenge. Traditionally, the concept of extracting patterns from data is similar with data mining. However, digging valuable information from large volume of data is far from simple data mining. This is because the volume of data and the velocity of data generation limit the efficient use of simple algorithms from data mining. Therefore, we summarize the key indicators of data-driven pattern prediction:

- **Accuracy.** The main indicator of a prediction model is its accuracy. It is awful to imagine that making prediction is just like guessing randomly. The accuracy of models would no doubt influence the use of patterns and leading to misleading behaviors and incurring great loss. Thus, pattern prediction without accuracy is useless to practical applications.
- **Efficiency.** Due to the large volume of generated data, if the models cannot treat data rapidly, the prediction would lose its timelessness. For example, if an application wants to decide how to conduct resource management for submitted requests, then it is unsuitable to queue requests for long time to conduct online prediction. We should mention here that the rapid response of prediction is the requirement for online scenario, not the offline data training progress. This is because it is worthy spending adequate time on offline training in order to have quick and accurate response for those online demands.



Data-Driven Pattern Prediction, Fig. 1 The population of Manhattan, hour by hour (2018). (a) Monday, 8:00. (b) Monday, 12:00. (c) Monday, 16:00. (d) Monday, 20:00. (e) Tuesday, 8:00. (f) Tuesday, 12:00. (g) Tuesday, 16:00. (h) Tuesday, 20:00

- **Scalability.** Under realistic scenarios, the volume of datasets and the dimension of features are extremely high. How to design an algorithm with high scalability to extract the patterns from big data is important. This is because it can help us find out the best suitable scenario to use such of these algorithms.

Fortunately, we can use the techniques from data mining as a guide to help design data-driven pattern prediction algorithms. Most of the techniques from machine learning, statistics, and database can be used to data-driven pattern prediction through some revisions. For example, we can use deep learning, reinforcement learning, and so on to capture the patterns for realistic applications. And at the same time, we should pay more attention on how to apply those approaches with high accuracy, efficiency, and scalability. Basically, data-driven pattern prediction has following steps:

- **Data Preprocessing.** The main idea behind for data preprocessing is to form the raw data into regular structures and at the same time, to treat the missing data for use. This is because the raw data collected from real world may have redundant information or missing fields.
- **Pattern Extraction.** This step focuses on extracting valuable patterns or conclusions from preprocessed datasets by using the techniques from machine learning, statistics, and database. We should mention here that the efficiency and accuracy are of great concern. This is because misleading conclusions or delayed news have less value for realistic scenarios.
- **Prediction based on Patterns.** This step mainly uses extracted patterns or conclusions from data to guide improving performance or solving meaningful problems, which is the key step to show the value of data.

Cross-References

- [Big Data Networks](#)
- [Data Mining](#)

References

- Big Data (2016). https://en.wikipedia.org/wiki/Big_data
- Lena long (2013). What Happens in an Internet Minute. <http://www.dailyinfographic.com/what-happens-in-an-internet-minute-infographic>
- Mta Subway Data (2018). <https://spatialityblog.com/2010/07/08/mta-gis-data-update/>
- Population of Manhattan, Hour by Hour (2018). <https://untappedcities.com/2018/05/15/>
- Turnstile Data of Metropolitan Transportation Authority (2018). <http://web.mta.info/developers/turnstile.html>
- Zhang H, Ananthanarayanan G, Bodik P, Philipose M, Bahl P, Freedman MJ (2017) Live video analytics at scale with approximation and delay-tolerance. In: NSDI, vol 9, p 1

Data-Driven QoE Measurement

Xiaoming He¹ and Kun Wang²

¹College of IoT, Nanjing University of Posts and Telecommunications, Nanjing, China

²Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA

Synonyms

[Data-access QoE management](#); [Data-aware QoE management](#); [Data-driven QoE analysis](#); [Data-driven QoE assessment](#); [Data-driven QoE improvement](#); [Data-driven QoE management](#); [Data-driven QoE optimization](#); [Data-driven QoE prediction](#); [Data-driven QoE visualization](#); [Data-oriented QoE management](#)

Data-Driven QoE Analysis

- [Data-Driven QoE Measurement](#)

Data-Driven QoE Assessment

- [Data-Driven QoE Measurement](#)

Data-Driven QoE Improvement

- [Data-Driven QoE Measurement](#)

Data-Driven QoE Management

- [Data-Driven QoE Measurement](#)

Definition

Recently, the strict definition of data-driven QoE measurement has not been given. The International Telecommunication Union (ITU-T) has defined the QoE concept as *the entire thing of availability of services subjectively perceived by end users*. The definition of QoE by the European Qualinet is the degree of satisfaction or annoyance of the end users of services because the utility and/or the expectations regarding services are based on end-user attitudes and current service situations (He et al. 2018). Furthermore, data-driven QoE measurement can be defined from the perspective of objective-based or subjective-based metrics. Generally speaking, data-driven QoE measurement is the end-user-centric metric that is used to measure end-user acceptance or satisfaction of data services or applications. In addition, data-driven QoE measurement is the extended version of data-driven QoS measurement used for judging delay-sensitive data from the end-user perspective (Wang et al. 2018). In summary, the common understanding of data-driven QoE

measurement is as follows: data-driven QoE measurement is a new measurement for data services which is based on the interactions between end users and services.

Historical Background

The data-driven QoE concept is a well-known measurement mechanism for determining the overall perception of the data-driven QoS, i.e., the evaluation of data-driven QoS as experienced by end users. Recently, many data-driven QoS-based methods have been developed to optimize the efficiency and performance of the environment, as proposed by the authors in Fang et al. (2016). Even though the parameters of data-driven QoS offer good objective measurement criteria, they cannot directly determine the quality of end-user perceptions. Therefore, both academic and industry researchers have shifted their attention from data-driven QoS parameters like jitter, throughput, packet loss, and delay to the concept of data-driven QoE. Data-driven QoE, in contrast, can refer to both the performance and efficiency of services as measured by data-driven QoS, as well as the subjective opinions of end users. Therefore, data-driven QoE is more suitable with respect to end users in a form of measurement than is data-driven QoS.

In fact, data-driven QoE analysis is paid close attention to the data-driven QoS metrics, i.e., bitrate, rebuffering, and delay. Data-driven QoE analysis models have appeared of late due to availability of large-scale data accompanied by the promising popularity of traffic and/or streaming around the Internet. In particular, video streaming/traffic has the availability of large-scale data in order to analyze QoE around the Internet (Wang et al. 2019). Later on, some extensions to the data-driven QoE analysis model have been studied. QoE analysis model with linear regression and correlation has been proposed in Du et al. (2018). The Kendall correlation coefficient can be selected to calculate the correlations between each metrics of data-driven QoS and data-driven QoE. Then the collection of information can help to get a

deeper understanding of the relationship between data-driven QoS and QoE via refining how the intellectual of uncertainty of the data-driven QoE metrics increases a certain data-driven QoS metric. Finally, linear regression-based curve fitting is the application to the data-driven pairs of QoS-QoE. By observing the data-driven QoS-QoE curves, the relationship is not linear in the whole environments. Additionally, a decision tree-based QoE prediction model is developed in Wang et al. (2016) to eliminate defects of correlation and linear regression analysis.

Also, data-driven QoE management has been studied. For example, a data-driven management architecture for personalized QoE is proposed to assess QoE from user perspective involving the subjective characteristics of end users (Wang et al. 2016). Then, a reinforcement learning-based method has been used for QoE-based spectrum handoff management (Xu et al. 2018). Data-aware QoE-QoS management method based on principle component analysis proposed to balance QoE-QoS management (Usman et al. 2018).

Lastly, several data-driven QoE manners, e.g., evaluation, and visualization have been proposed to enable data-driven QoE optimization in Jiang et al. (2017). Content providers can get amazing profits from data-driven platform via transferring strategies of load balance based on QoE server statement visualization and monitoring (Wang et al. 2016).

Key Applications

The common understanding of data-driven QoE is a novel measurement technique for applications or services and is confirmed based on the quality of whole environments and the experience of end users. Data-driven QoE measurement has been applied to a variety of scenarios.

Video streaming The flexibility is used to stream interesting content from CDNs and/or multiple servers in different bitrates by VoD offers that already put real-time visibility into video quality by the instrumentation of client (Zhang et al. 2017). In contrast to the

traditional methods, current work has displayed the operation on a single-session basis; data-driven methods for CDN selection and bitrate can enhance the quality of video (e.g., 50 percent less buffer in time) (Jiang et al. 2016a). Similarly, the overlay path between edge servers and sources is the dependence of live streaming QoE (e.g., ESPN, twitch.tv). Once these overlay paths are the choices to tackle traffic and the fluctuation of quality dynamically, there could be development space for enhancement. Existing works have been introduced to leverage video quality in the following.

- **Design of media player buffer:** The design of media player buffer is crucial adjective because the event of rebuffering has a vital impact on QoE of end users. The size of buffer will affect the delay and time of rebuffering. The delay will be longer since more and more data can be downloaded before the player plays if the buffer size is large. During the playing state, vice versa, however, fewer events of rebuffering can happen. Additionally, He et al. (2018) has been shown that most of the downloaded data in the buffer is not useful due to many end users quitting before the completions of video.
- **Congestion control:** Conventional TCP protocol applied to video traffic can result in long delay owing to the following reasons: (1) considering the TCP protocol, a lost packet will be retransferred when it is received successfully, leading to a long delay and get a poor QoE; (2) the additive increase multiplicative decrease (AIMD) algorithm results in fluctuated throughput around the time, which will increase the delay and result in end-user dissatisfaction; and (3) the congestion control is QoS-based, whereas the video is QoE-based. Sterca et al. (2016) proposed the Media-TCP to design a mechanism of congestion control video-friendly for the TCP protocol, optimizing the congestion window size, which maximize the long-term QoE. The distortion delay and influence deadline of

each packet are considered to offer different services or applications for different packet classes. Media-TCP is shown to improve the PSNR around methods of conventional TCP congestion control. Unfortunately, Wang et al. (2017b) proposed Media-TCP, which is still QoS-based, a MOS-based congestion control for multimedia transmission. The value of MOS is predicted in real-time manner via the Lync system of Microsoft on the basis of quantitative measurements (i.e., packet loss, bit errors, packet delay, and jitter). Then QoE-based adaption of congestion window is formulated as a partially observable Markov decision process (POMDP) and is solved by the algorithm of online learning. Another method is the use of protocol such as video-friendly application, Dynamic Adaptive Streaming over HTTP (DASH), mitigating the delay problem in video transmission and not changing the TCP protocol.

Internet telephony Recently, applications of VoIP (i.e., Skype and Hangouts) are useful for managed relays, optimizing performances between network clients. In contrast to any cast-based relay selection, recent work has been shown that a data-driven algorithm of relay selection can reduce the number of poor-quality calls (i.e., 42 percent fewer calls with over 1.2 percent packet loss) (Canbaz et al. 2017). Thereby, the satisfaction of QoE can be given by the proposed data-driven algorithm in the Internet telephony.

File sharing The flexibility of file-sharing applications (e.g., Dropbox) allows each client to catch files from a chosen data center or server. We can enhance the QoE for these services or applications potentially by the use of data-driven methods to predict the throughput between a server and client (Chen et al. 2015).

Web services With the applications of social network, we will attempt to optimize server selection. The goal is to get information of the co-

location. Recent work has shown that optimal caching and server selection can reduce query time of Facebook to 50 percent (Du et al. 2018). Online services (e.g., search) can also benefit from data-driven techniques. For example, compared to anycast, edge sever selection of data can reduce search latency by 60 percent serving $2\times$ more queries. Therefore, using data-driven technology in Web services is beneficial to the reduction of latency and, furthermore, brings better QoE performance.

Cross-References

- [QoE Assessment Models](#)
- [Data-Driven QoE Optimization](#)

References

- Canbaz M, Thom J, Gunes MH (2017) Comparative analysis of internet topology data sets. In: Proceedings of 2017 IEEE conference on computer communications workshops (INFOCOM WKSHPS)
- Chen K, Shen H (2015) Maximizing P2P file access availability in mobile ad hoc networks though replication for efficient file sharing. *IEEE Trans Comput* 64(4):1029–1042
- Du M, Wang K, Xia Z, Zhang Y (2018) Differential privacy preserving of training model in wireless big data with edge computing. *IEEE Trans Big Data*. <https://doi.org/10.1109/TBDDATA.2018.2829886>
- Fang Q, Wang J, Gong Q (2016) QoS-driven power management of data centers via model predictive control. *IEEE Trans Auto Sci Eng* 13(4):1557–1566
- He X, Wang K, Huang H, Miyazaki T, Wang Y, Guo S (2018) Green resource allocation based on deep reinforcement learning in content-centric IoT. *IEEE Trans Emerg Top Comput*. <https://doi.org/10.1109/TETC.2018.2805718>
- He X, Wang K, Huang H, Miyazaki T, Wang Y, Sun Y (2018) QoE-driven joint resource allocation for content delivery in fog computing environment. In: Proceedings of ICC.
- Jiang J, Sekar V, Milner H, Shepherd D, Stoica I, Zhang H (2016a) CFA: a practical prediction system for video QoE optimization. In: Proceedings of 13th USENIX symposium on networked systems design and implementation (NSDI'16)
- Jiang J, Sun S, Sekar V, Zhang H (2017) Pytheas: enabling data-driven quality of experience optimization using group-based exploration-exploitation. In: Proceedings of 14th USENIX symposium on networked systems design and implementation (NSDI'17)
- Purohit L, Kumar S (2018) A classification based web service selection approach. *IEEE Trans Serv Comput*. <https://doi.org/10.1109/TSC.2018.2805352>
- Sterca A, Hellwagner H, Florian Boian F, Vancea A (2016) Media-friendly and TCP-friendly rate control protocols for multimedia streaming. *IEEE Trans Circ Syst Video Technol* 26(8):1516–1531
- Usman M, Yang N, Jan M, He X, Xu M, Lam k (2018) A joint framework for QoS and QoE for video transmission over wireless multimedia sensor networks. *IEEE Trans Mob Comput* 17(4):746–759
- Wang K, Gao H, Xu X, Jiang J, Yue D (2016) An energy-efficient reliable data transmission scheme for complex environmental monitoring in underwater acoustic sensor networks. *IEEE Sens J* 16(11):4051–4062
- Wang K, Mi J, Xu C, Zhu Q, Shu L, Deng DJ (2016) Real-time load reduction in multimedia big data for mobile Internet. *ACM Trans Mult Comput Commun App* 12(5):76
- Wang K, Zhuo L, Shao Y, Yue D, Tsang KF (2016) Towards distributed data processing on intelligent leak-points prediction in petrochemical industries. *IEEE Trans Ind Informatics* 12(6):2091–2102
- Wang Y, He D, Ding L, Zhang W, Li W, Wu Y, Liu N, Wang Y (2017b) Media transmission by cooperation of cellular network and broadcasting network. *IEEE Trans Broad* 63(3):571–576
- Wang K, Zhou Q, Guo S, Luo J (2018) Cluster frameworks for efficient scheduling and resource allocation in data center networks: a survey. *IEEE Commun Surv Tutor* 20(4):3560–3580
- Wang Y, Wang* K, Huang H, Miyazaki T, Guo S (2019) Traffic and computation co-offloading with reinforcement learning in fog computing for industrial applications. *IEEE Transactions on Industrial Informatics*. *IEEE Trans Ind Informatics* 15(2):976–986
- Xu C, Wang K, Li L, Xia R, Guo S, Guo M (2018) Renewable energy-aware big data analytics in geo-distributed data centers with reinforcement learning. *IEEE Trans Netw Sci Eng*. <https://doi.org/10.1109/TNSE.2018.2813333>
- Zhang J, Wang Z, Wang K, Guo S, Guo M (2017) Improving power efficiency for online video streaming service: a self-adaptive approach. *IEEE Trans Sust Comput*. <https://doi.org/10.1109/TSUSC.2017.2739798>

Data-Driven QoE Optimization

- [Data-Driven QoE Measurement](#)

Data-Driven QoE Prediction

► [Data-Driven QoE Measurement](#)

Data-Driven QoE Visualization

► [Data-Driven QoE Measurement](#)

Data-Driven Resource Allocation

He Li

Department of Information and Electronic Engineering, Muroran Institute of Technology, Muroran, Japan

Synonyms

[Big data computing](#); [Data-driven scheduling](#); [Distributed computing](#)

Definition

Data-driven resource allocation is a methodology that resource allocation in computer systems is mainly according to data storage and processing. In data-driven resource allocation, resources need to be allocated for supporting the data-flow between storage and processors. Since big data application usually needs data-intensive computing, it is very important to allocate resource for massive data storage and transfer.

Historical Background

Resource allocation is an important problem in modern computing systems (Dong et al.

2008). When a task is executed in a general computing system, it is necessary to allocate enough resources for it to normally work. Usually, resources in computing systems include memory space, storage space, processor time, network bandwidth, and so on. For a multitasking computing system, it needs a dedicated resource allocation strategy to finish tasks with limited resources.

There are many resource allocation strategies in different computing systems. In early multitasking computers, the operating systems allocate fixed resources for different tasks (Vasudevan et al. 2015). For example, in the oldest round-robin scheduling, the task scheduler will assign a fixed number of processor time slices to each task (Hahne 1991). However, since required resources of tasks are different, fixed resource allocation strategy is equitable but not efficient.

Another methodology of resource allocation is the on-demand strategy that resources are allocated to tasks according to their requests (Albonesi 1999). In memory space allocation, operating systems usually assign required memory of each task. Thus, the maximum number of executing tasks in a computing system is usually limited by the total memory space (Avisar et al. 2002).

In a distributed computing system, the resource allocation problem becomes more complex because of the communication between distributed servers (Kurose and Simha 1989). Usually, since most distributed computing systems are built on commercial hardware, the communication bandwidth is more limited than processor time and memory space. The allocation strategy of network resources affects the performance of the total distributed systems (Li et al. 2014).

Therefore, in network bandwidth allocation for distributed systems, it needs to understand the traffic of each task and then allocate the appropriate bandwidth to tasks. Since most network traffic between servers depends on the data-flows of each task, a data-driven network resource allocation plays an important role in distributed computing systems (Li et al. 2015b).

Data-intensive computing such as big data processing and multimedia computing becomes more and more important in recent years (Wu et al. 2016). For data-intensive computing tasks, the storage resources are more important than processor resources. Due to slow storage hardware, high-speed memory allocation for data-intensive tasks is much harder than compute-intensive tasks (Li et al. 2015a). Allocating resource for data-intensive tasks becomes another important area of data-driven resource allocation.

Data-driven resource allocation is also very important for stream computing that processes large data streams from real-time sources. Since data-flows of different stream computing tasks are usually different, it needs a suitable resource allocation paradigm to allocate resources accordingly and to predict data-flow of each task.

Researchers begin to apply data-driven resource allocation paradigm in some new problems. For example, in training deep learning models, a large amount of data will be transferred from storage devices or memory to some specific computing devices such as graphics processing units (GPUs) or application-specific integrated circuit (ASIC) chips. Therefore, data-driven resource allocation strategies are applied to optimize the resource allocation in training varying deep learning models.

Key Applications

According to the importances discussed above, data-driven resource allocation can be applied in many applications such as big data computing, cloud computing, edge computing, deep learning, and Internet-of-Things (IoT).

1. Big data computing

In big data computing, data-driven resource allocation plays a very important role for the data processing efficiency. For example, in traditional MapReduce computing model, the resource

allocation needs to follow the data-flows due to the heavy traffic and IO from mapping to reducing phase.

2. Cloud computing

Cloud offers computing resources for different applications with the pay-as-you-go model. For cloud providers, it is still a difficult problem to allocate resources to different applications with quality-of-service (QoS) guarantees. There are more and more data-intensive applications being deployed on cloud platforms in recent years. Due to the high concentration of the IO model applied in cloud computing, efficient data-driven resource allocations are essential to QoS-guaranteed cloud services.

3. Edge computing

Edge computing can off-load computing or storage loads from cloud servers to the edge. Usually, due to the high mobility of edge devices, it is very important to allocate resources according to the mobile data-flows between edge nodes and devices. Meanwhile, it also needs a data-driven resource allocation for different services deployed in the edge since most edge devices or nodes have limited storage space and computational capacity.

4. Deep learning

In deep learning tasks, especially in training deep learning models, massive datasets are transferred from storage devices to the learning networks. Since the learning networks become more and more complex, it needs distributed learning systems to reduce the training time. Meanwhile, the intermediate data are shuffled between different computational devices in training tasks. Therefore, with the rapid development of deep learning technologies, it is necessary to work out deep learning-oriented data-driven resource allocation strategies according to the massive data-flows in learning systems.

5. IoT

The IoT technology connects various devices such as home appliances, vehicles, or industrial machines embedded with processors, sensors, and network connections. In general, in IoT applications, a lot of sensing information is transferred from IoT devices to servers for further processing. It needs to allocate resources according to data-flows from devices to servers in building IoT applications. The challenge of data-driven resource allocation in IoT is to allocate resources for massive IoT devices since the small size and large scale of data-flows in IoT are different from general computing environments.

Cross-References

- [Data-Driven Edge Computing](#)
- [Resource Allocation](#)
- [Resource Management](#)

References

- Albonesi DH (1999) Selective cache ways: on-demand cache resource allocation. In: MICRO-32. Proceedings of the 32nd annual ACM/IEEE international symposium on microarchitecture, pp 248–259. <https://doi.org/10.1109/MICRO.1999.809463>
- Avissar O, Barua R, Stewart D (2002) An optimal memory allocation scheme for scratch-pad-based embedded systems. *ACM Trans Embed Comput Syst* 1(1):6–26. <https://doi.org/10.1145/581888.581891>
- Dong M, Guo M, Zheng L, Guo S (2008) Performance analysis of resource allocation algorithms using cache technology for pervasive computing system. In: 2008 the 9th international conference for young computer scientists, pp 671–676. <https://doi.org/10.1109/ICYCS.2008.527>
- Hahne EL (1991) Round-robin scheduling for max-min fairness in data networks. *IEEE J Sel Areas Commun* 9(7):1024–1039. <https://doi.org/10.1109/49.103550>
- Kurose JF, Simha R (1989) A microeconomic approach to optimal resource allocation in distributed computer systems. *IEEE Trans Comput* 38(5):705–717. <https://doi.org/10.1109/12.24272>
- Li Z, Guo S, Zeng D, Barnawi A, Stojmenovic I (2014) Joint resource allocation for max-min throughput in multicell networks. *IEEE Trans Veh Technol* 63(9):4546–4559. <https://doi.org/10.1109/TVT.2014.2317235>
- Li H, Dong M, Liao X, Jin H (2015a) Deduplication-based energy efficient storage system in cloud environment. *Comput J* 58(6):1373–1383. <https://doi.org/10.1093/comjnl/bxu122>
- Li H, Guo S, Wu C, Li J (2015b) Fdrc: flow-driven rule caching optimization in software defined networking. In: 2015 IEEE international conference on communications (ICC), pp 5777–5782. <https://doi.org/10.1109/ICC.2015.7249243>
- Vasudevan SK, Vasudevan S, Velusamy K, Muralidharan S (2015) Modern operating systems, 1st edn. I.K. International Publishing House, New Delhi
- Wu J, Guo S, Li J, Zeng D (2016) Big data meet green challenges: big data toward green applications. *IEEE Syst J* 10(3):888–900. <https://doi.org/10.1109/JSYST.2016.2550530>

Data-Driven Scheduling

- [Data-Driven Resource Allocation](#)

Data-Driven Security

- Guangquan Xu³, Shuhan Huang³, Bingjiang Guo³, Ning Zhang^{1,2,3}, Wenjuan Zhou³, Mengdi Liu³, Xianjiao Zeng³, Ran Wang^{3,4,5}, and Longpeng Li³
- ¹Peng Cheng Laboratory, Shenzhen, China
- ²Texas A&M University-Corpus Christi, Corpus Christi, TX, USA
- ³College of Intelligence and Computing, Tianjin University, Tianjin, China
- ⁴Nanjing University of Aeronautics and Astronautics, Nanjing, China
- ⁵Department of Information and Communication Engineering, Tongji University, Shanghai, China

Synonyms

[Data specified security](#)

Definitions

Data-driven security means that in an environment where all decisions and processes are determined by data, security professionals promote the

behavior of security programs by balancing the following factors: selecting and deploying effective security measures, working within budget limits, and reducing liability exposure.

Historical Background

It has been long time since researchers began to learn from data, but data-driven security was just proposed in twenty-first century (Zhang et al. 2011). According to the way of data utilization, data analysis and processing technology in solving network security problems has gone through three phases so far (Jacobs and Rudis 2014). The primary application area of the first phase is intrusion detection. In particular, according to the difference of data source, intrusion detection technology is divided into host intrusion detection and intrusion detection based on network (Huang 2018). According to the method of data analysis (Comaniciu 2003), it can be divided into the detection based on misuse (feature) and the detection based on anomaly (Tu and Zhu 2002; Tu 2002). Meanwhile, data mining and machine learning began to be applied to test and defend the invasion of network and system (De Vries et al. 2010). Data mining or machine learning essentially establishes a classifier model so as to output the different classification results by identifying the input data (Nivre et al. 2007; Kim et al. 2018; Matusik 2003). The representative product form of the second phase is SIEM (Miller et al. 2010). To unify security information and event management, the first step is to collect the alarm information of different intrusion detection systems and products for correlation analysis (Nath et al. 2010). Then information with operational capability is excavated, so that the security experts, operators, and managers can take timely and effective response measures to deal with more complex network security situational changes. As the data analysis technology improving, the data-driven security begins to attract network and information security specialists to study on it. Nowadays, data-driven security has been applied in many aspects, such as finance, health service, industry, education, and so on. And the

data-driven security technology is playing a more and more significant role in information security defense area, including instruction detection, exploit and analyze vulnerability, fraud detection, phishing attack detection, and spam detection. Due to the qualitative change of underlying data and computing ability, now the data-driven security products can also make an analysis on source code, binaries, and huge amounts of unstructured data such as audio and video and image. The time span of correlation analysis is no longer limited to the hour or day level, but it is increased to the level of the month or year (Fan et al. 2014).

Foundations

Data-driven security is of great significance to cyber security, especially in the era of big data. Today, data is a treasure filled with information instead of binary numbers. Similarly, the advantages of big data are also used in the field of cyber security in business and scientific research. Including security audit, cyber security situational awareness, log analytics, network attack tracing, and so forth, related security technologies have been further improved in intelligence. In this situation, complex security issues can be quickly resolved. However, it will also face new challenges in practice as follows:

1. **Big Data Processing.** Big data is often accompanied by heterogeneity, high dimensionality, heterogeneity, storage bottleneck, and noise accumulation (Fan et al. 2014). Therefore, if there are no appropriate tools and methods for data collection, processing, and storage, then the security analysis work will be extremely difficult (Si et al. 2011).
2. **Programming and Security Technology for Big Data.** People generate data from various behaviors on the web, such as web browsing, upload and download, instant messaging, live broadcast (Zhang et al. 2005), etc. The required analytical capability far exceeds that of traditional safety methods and tools. Thus, it is necessary to find programming and

security technologies that adapt to big data (Jacobs and Rudis 2014).

3. **Methodology for Mining Security Issues.** Data itself is worthless, and it just consists of a large number of characters in various ways. We can use the value of data only when we refine the data into effective information and further sublimate them into knowledge. Therefore, statistical knowledge and techniques as data mining methodology are indispensable for cyber security.
4. **Information Overload.** It easily leads to information overload problems during data mining, especially when the data is getting bigger. Although there is a lot of effective information, we cannot make good use of it. To this end, security visualization technology is urgently needed to solve this problem.

At present, most researches focus on solving the above problems. Therefore, the researches are mainly divided into the following four categories:

1. Data Management

Data Management is a collection of practices, concepts, procedures, and processes that allow for an organization to gain control of its data resources. It is involved with the entire lifecycle of a given data asset from its original creation point to its final retirement. And the Data Management Body of Knowledge (DMBOK) (Mosley et al. 2010) refers to data management as “The planning, execution, and oversight of policies, practices and projects that acquire, control, protect, deliver, and enhance the value of data and information assets.”

Data operations management is a central function of the entire data management framework. And it is defined as the development, maintenance, and support of structured data, which can maximize data resources’ value to the enterprise (Han et al. 1993).

While the above definition says structured data, the definition needs to be expanded to include semi-structured and unstructured data that now exists due to the rise of such contemporary data management trends like big

data, cloud computing, machine learning, data lakes, and the Internet of Things.

The practice of data management includes an extensive list of associated and related topics which span the entire process of managing and leveraging data at all levels. A short list of these includes:

- **Data Governance and Data Stewardship**

Data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data.

Data stewardship is the management and oversight of an organization’s data assets to help provide business users with high-quality data that is easily accessible in a consistent manner.

- **Data Architecture**

Data architecture is composed of models, policies, rules, or standards that govern which data is collected and how it is stored, arranged, integrated, and put to use in data systems and in organizations. Data is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.

- **Data Modeling**

Data modeling is the process of documenting a complex software system design as an easily understood diagram, using text and symbols to represent the way data needs to flow. The diagram can be used to ensure efficient use of data, as a blueprint for the construction of new software, or for reengineering a legacy application.

- **Data Quality Management**

Data quality management (DQM) refers to a business principle that requires a combination of the right people, processes, and technologies all with the common goal of improving the measures of data quality that matter most to an enterprise organization. That last part is important: the ultimate purpose of DQM is not just to improve data quality for the sake of having high-quality data but rather to achieve the business outcomes that depend upon high-quality data.

The big one is customer relationship management or CRM.

- **Master and Reference Data Management**

Master data management (MDM) is a method used to define and manage the critical data of an organization to provide, with data integration, a single point of reference. The data that is mastered may include reference data – the set of permissible values – and the analytical data that supports decision-making.

Reference data management is a process that drives efficiency across the market by maintaining high data quality standards. It provides full terms and conditions data, including ratings and descriptive content, for each asset types.

- **Data Warehousing**

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision-making.

- **Big Data**

Big data is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

- **Business Intelligence and Analytics**

Business intelligence (BI) is a technology-driven process for analyzing data and presenting actionable information to

help executives, managers, and other corporate end users make informed business decisions.

Business analytics (BA) is the practice of iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies committed to data-driven decision-making.

- **Metadata Management**

Metadata management is the end-to-end process and governance framework for creating, controlling, enhancing, attributing, defining, and managing a metadata schema, model, or other structured aggregation system, either independently or within a repository and the associated supporting processes (often to enable the management of content). For web-based systems, URLs, images, video, etc. may be referenced from a triple table of object, attribute, and value.

- **Data Security Management**

Data security management is a way to maintain the integrity of data and to make sure that the data is not accessible by unauthorized parties or susceptible person to corrupt it. Data security is put in place to ensure privacy in addition or protecting this data. Data itself is a raw form of information that is stored on network servers, on possible personal computers, and in the form of columns and rows. This data can be anything from personal files to intellectual property and even top-secret information. Data can be considered as anything that can be understood and interpreted by humans. Because of the growing use of Internet, there is an emphasis on protecting personal or company data, and this protection is known as data security which can be acquired by using specific software solutions or hardware mechanisms.

- **Content Management**

Content management is a set of processes and technologies that supports the collection, managing, and publishing of

information in any form or medium. When stored and accessed via computers, this information may be more specifically referred to as digital content or simply as content.

A well-developed data management program within an organization has the ability to positively affect change around the administration and use of data assets across all levels, departments, and lines of business. The benefits of data management include improvements of operations management, more effective marketing and sales, better regulation and compliance controls, enhanced security and privacy, reduction of risk across the board, faster application and system development, improved decision-making and reporting, sustained business growth, business and technical alignment, automated and/or streamlined operations, greater collaboration, revenue growth, more consistency across all enterprise processes, and numerous others.

2. Programming and Security Technology

• Programming Technology

Data may come in a wide range of formats, states, and quality. It may be embedded in unstructured or semi-structured log files. Somehow, we must find a way to collect, coax, combine, and massage what we are given into a format that supports further analysis. And learning even basic programming skills opens up a whole range of possibilities when working with data. It frees you to accept multiple forms of data and manipulate it into whatever formats work best with the analysis software you have.

Modern programming language supports basic data manipulation tasks, but scripting languages such as Python and R appear to be used slightly more often in data analysis than their compiled counterparts (Java and C). However, the programming language is somewhat irrelevant. The end results (and a happy analyst) are more important than picking any “best” language. Whatever gets the

job done with the least amount of effort is the best language to use. We generally flip between Python (pandas) and R for cleaning and converting data (or perhaps some Perl if we’re feeling nostalgic) and then R or pandas for the analysis and visualization. In addition, learning web-centric languages like HTML, CSS, and JavaScript will help create interactive visualizations for the web, but web languages are not typically involved in the preparation and analysis of data.

• Security Technology

– Security Audit

It is a systematic evaluation of the security of a company’s information system by measuring how well it conforms to a set of established criteria. A thorough audit typically assesses the security of the system’s physical configuration and environment, software, information handling processes, and user practices. And in the wake of relevant legislations, security audits are often used to determine the regulatory compliance of a system (Rouse 2018).

– Cyber Security Situational Awareness

It refers to an understanding of the cyber threat environment, including how it operates, its associated risks and impacts, and the mitigation measures of these risks. A keen appreciation of the threat landscape can help businesses and individuals to better understand the vulnerabilities so as to adopt early appropriate mitigation measures. We encourage industry players and organizations to have active participation in information-sharing arrangement and collaboration with trusted stakeholders locally and globally for timely and relevant cyber threat intelligence.

– Log Analytics

With data breaches becoming more frequent and sophisticated, protecting customer information and intellectual property is of paramount importance.

The security log analytics solution will enable security teams to gain comprehensive visibility into their environment and detect anomalous behavior as quickly as possible.

Detect anomalous behavior: Early detection of advanced persistent threats and unknown threats.

Minimize exposure: Avoid fines, lawsuits, loss of business and negative PR.

Take quick action: React fast on any abnormal or malicious activity from internal and external actors.

3. The intersection of data-driven security and statistics

Data-driven security is based on the big data, by building prediction models and algorithms based on statistics to determine potential threats and risks. Jeyakumar et al. (2016) constructed an algorithm that used statistics methods to classify the system malware and predict potential malicious attacks by analyzing the processor and memory information of a large number of sample hosts. Interpretation and prediction of classical statistics, also known as feature extraction of data in the field of machine learning, are the core tasks of data analysis, such as ports, protocols, etc., are characteristics of network traffic security analysis. And we need to select the global optimal feature set through the method of optimal subset and stepwise comparison based on statistical principles. At the same time, the work of data-driven security predictions needs to find potential relationships through statistics based on supervised learning and unsupervised learning methods (Cimiano and Volker 2005).

4. Security Visualization

The most effective way to recognize human beings is through visual perception. Therefore, data visualization is used to provide the feedback of analysis for data-driven security.

- Security visualization can identify potential patterns. Some patterns that exist in a single variable or relationships among multiple variables cannot be identified by statistical methods for the reason

that visualization techniques are based on physiologic saccades, which give important information with more visible visual features, making it easy to recognize the intrinsic relationships among variables.

- Security visualization can rapidly complete information transmission. Classical statistical methods such as variance and mean value lose the subtle characteristics of data transfer, while visualization minimizes the loss of data details and enables that communicating millions of point information can be done in a few seconds, facilitating macro and micro safety analysis of data.

Key Applications

Data-driven security is often applied to instruction detection, exploit and analyze vulnerability, fraud detection, phishing attack detection, spam detection, security audit, cyber security situational awareness, log analytics, network attack tracing, and so forth.

References

- Cimiano P, Volker J (2005) A framework for ontology learning and data-driven change discovery. *Lect Notes Comput Sci* 3513:227–238
- Comaniciu D (2003) An algorithm for data-driven bandwidth selection. *IEEE Trans Pattern Anal Mach Intell* 25(2):281–288
- De Vries SJ, Van Dijk M, Bonvin AM (2010) The haddock web server for data-driven biomolecular docking. *Nat protoc* 5(5):883
- Fan J, Han F, Liu H (2014) Challenges of big data analysis. *Natl Sci Rev* 1(2):293–314
- Han J, Cai Y, Cercone N (1993) Data-driven discovery of quantitative rules in relational databases. *IEEE Trans Knowl Data Eng* 5(1):29–40
- Huang W (2018) What to do data-driven security. <https://sec.cuc.edu.cn/huangwei/textbook/ns/ds-security/how.html>
- Jacobs J, Rudis B (2014) Data-driven security: analysis, visualization and dashboards. John Wiley & Sons Hoboken, NJ, America
- Jeyakumar V, Madani O, ParandehGheibi A, Yadav N (2016) Data driven data center network security. In: *Proceedings of the 2016 ACM on international workshop on security and privacy analytics*. ACM, pp 48–48

- Kim SH, Whitt W, Cha WC (2018) A data-driven model of an appointment-generated arrival process at an outpatient clinic. *Inform J Comput* 30(1):181–199
- Matusik W (2003) A data-driven reflectance model. PhD thesis, Massachusetts Institute of Technology
- Miller DR, Harris S, Harper A, VanDyke S, Blask C (2010) Security information and event management (SIEM) implementation (network pro library). McGraw Hill, New York, America
- Mosley M, Brackett MH, Earley S, Henderson D (2010) DAMA guide to the data management body of knowledge. Technics Publications
- Nath S, Vainstein K, Ouyé MM (2010) Method and system for securing digital assets using process-driven security policies. US Patent 7,703,140
- Nivre J, Hall J, Nilsson J, Chanev A, Eryigit G, Kübler S, Marinov S, Marsi E (2007) Maltparser: a language-independent system for data-driven dependency parsing. *Nat Lang Eng* 13(2):95–135
- Rouse M (2018) What is compliance. <http://search-datamanagement.techtarget.com/definition/compliance>
- Si XS, Wang W, Hu CH, Zhou DH (2011) Remaining useful life estimation—a review on the statistical data driven approaches. *Eur J Oper Res* 213(1):1–14
- Tu Z (2002) Image parsing by data-driven markov chain monte carlo. PhD thesis, The Ohio State University
- Tu Z, Zhu SC (2002) Image segmentation by data-driven markov chain monte carlo. *IEEE Trans Pattern Anal Mach Intell* 24(5):657–673
- Zhang J, Wang FY, Wang K, Lin WH, Xu X, Chen C (2011) Data-driven intelligent transportation systems: a survey. *IEEE Trans Intell Transp Syst* 12(4):1624–1639
- Zhang X, Liu J, Li B, Yum YS (2005) Coolstreaming/donet: a data-driven overlay network for peer-to-peer live media streaming. In: INFOCOM 2005. 24th annual joint conference of the IEEE computer and communications societies. Proceedings IEEE, vol 3. IEEE, pp 2102–2111

Data-Driven Service Provisioning

Huawei Huang¹ and Song Guo²

¹School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

²Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Synonyms

[Application of big data on service provisioning](#);
[Big data based service provisioning](#)

Definitions

Service provisioning is the process of configuring a service for customers. It usually associates with telecom industry and cloud infrastructure. The examples of service provisioning are shown as follows.

- When a customer submits a request to use a mobile network, system initializes and manages the customer's services, such as establishing control channels (Huang et al. 2016) and making update scheme for service provisioning due to user mobility (Huang and Guo 2017).
- To activate a low-earth-orbit satellite-based service for secure big data storage, the telecom company has to design and construct the infrastructure for customers in advance (Huang et al. 2018).
- Data centers or edge nodes launch a *Service Function Chaining* enabled machine to meet the customer's subscription, including the desired software and hardware resources, and the access authority in the agreed period (Huang et al. 2017b).
- Cloud infrastructure provider maintains a pool of instances. Customers are allowed to request these instances through application programming interfaces (API) (Xu et al. 2015; Huang et al. 2017a).

Background

Big data promises to reshape the knowledge, society, and economy. The technologies that are used to generate, collect, process, compute, and communicate big data have made significant progresses in the past few years. As an overwhelming volume of data is generated with ever-growing speed every day from various sources and applications, such as cloud services, Internet of Things (IoT), social network services, and other smart terminals, it becomes significant to design, deploy, and provision services that are more wise to enable the effective acquisition, storage, transformation, management, and utilization of such huge amount of data.

To alleviate the violations of *service-level objectives* caused by provisioning delays, inefficient reconfigurations, and inaccurate demand predictions, over-provisioning apparently increases operational costs and resource-utilization inefficiencies, particularly in large-scale systems. Via a data-driven service provisioning framework named AidOps, Lugones et al. (2017) argue that the orchestration system enables a higher service availability and lower over-provisioning costs.

Foundations

As mentioned by Chesbrough and Spohrer (2006), the growing knowledge of information and communication technology (ICT) has created opportunities to inspire the evolution of internet on the service relationships that can yield new value. Specifically, ICT has the following characteristics that can improve effectiveness and efficiency and enable innovativeness in the manners of (1) commoditizing noncore competency such as outsourcing, (2) improving

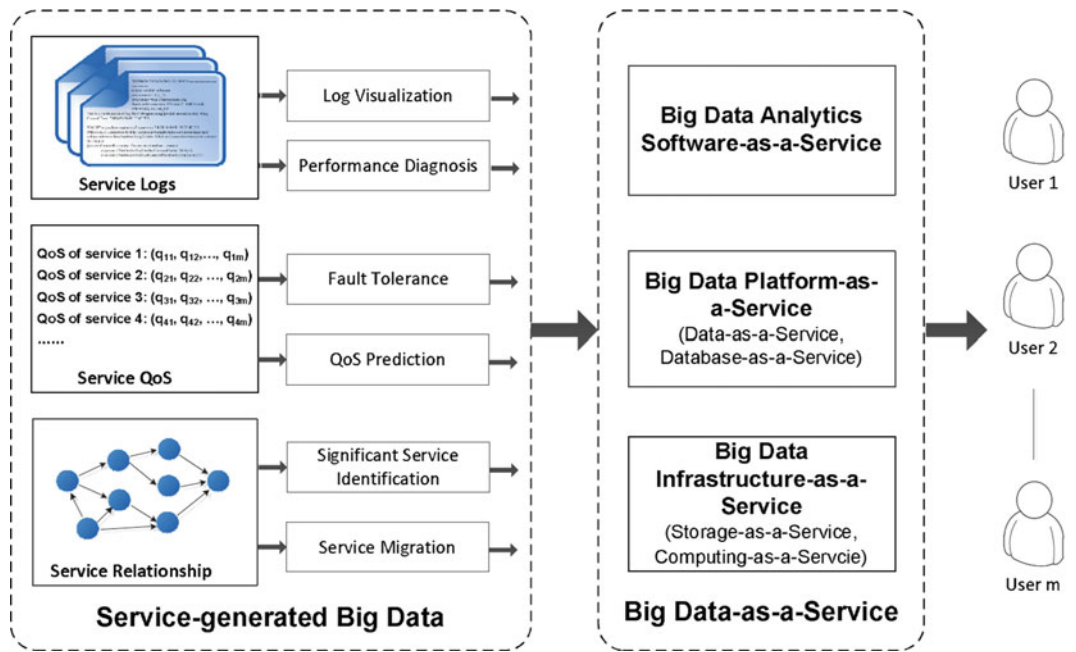
collaboration in business processes, (3) controlling the risks of information security violation, (4) coordinating service systems such as open innovation platforms, (5) improving customer-oriented service quality, etc. Especially, in the big data era, enormous value can be extracted from the service-oriented big data.

To fully explore the value of service-generated big data, developing efficient technologies that can fast process large amount of data is an open issue. On the other hand, easy access of big data and data analysis results are critical to both service providers and service users (Zheng et al. 2013).

Key Applications

In the following, some typical data-driven platforms/architectures for service provisioning are reviewed.

To provide the functionality of managing and analyzing service-generated big data, Zheng et al. (2013) proposed an infrastructure of data-driven service model. As shown in Fig. 1, a big



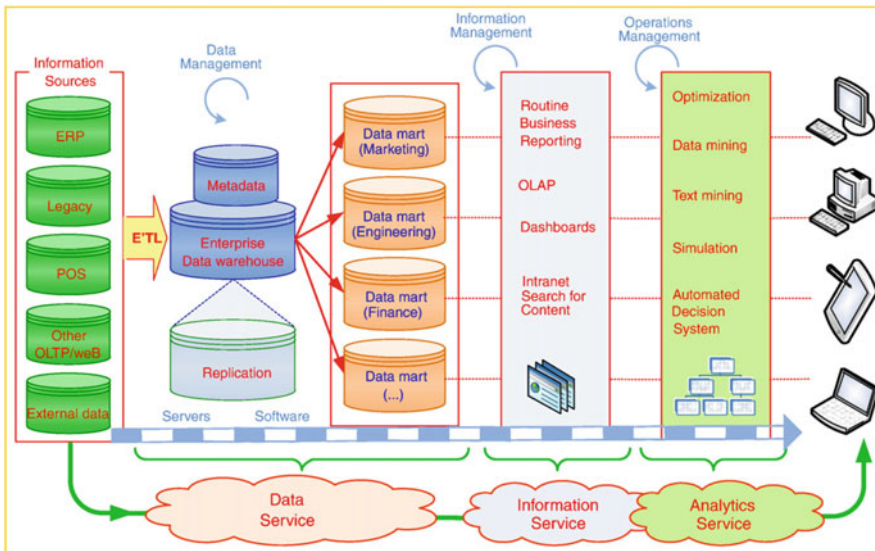
Data-Driven Service Provisioning, Fig. 1 Service-generated big data and big data-as-a-service as presented by Zheng et al. (2013)

data-as-a-service framework has been also employed to provide big data services and data analytics results to users. Under this framework, several concrete applications that exploit three types of service-generated big data are illustrating for performance enhancement. First, the service logs generated by service requests can be used to yield visualized log and conduct performance diagnosis. Second, based on the QoS information of each service, fault tolerance and QoS prediction can be performed by service providers. Finally, taking the service relationship into account, significant service identification and service migration are able to realize.

A framework for service-oriented decision support systems (DSS) has been investigated by Demirkan and Delen (2013). By focusing on the product-oriented environment and exploring engineering-related issues, a conceptual architecture of service-oriented DSS is shown in Fig. 2, which includes three service models, i.e., data-as-a-service, information-as-a-service, and analytics-as-a-service. In such a service-oriented DSS solution, the operational systems, data warehouses, online analytics processing, and

end-user components can be provided to users as services either separately or in the bundled manner.

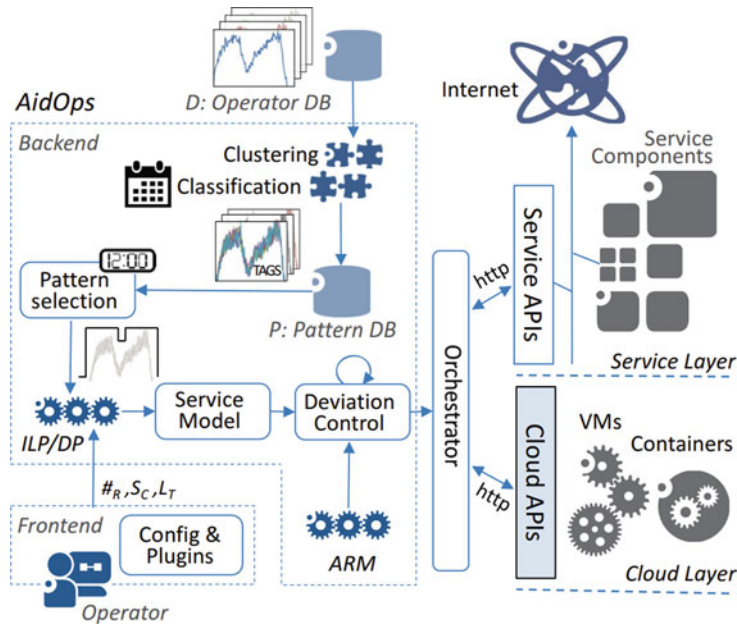
As mentioned previously, a systematic approach named AidOps for capacity provisioning of services in cloud is proposed by Lugones et al. (2017). Instead of on-demand elasticity, AidOps enables service providers to specify a desirable number of reconfigurations (i.e., scaling in/out) as an input, and it then calculates the capacity configurations required to serve the user demand optimally. Figure 3 illustrates the architecture design of AidOps, including system components and interfaces to the cloud and service layers. This system is consisted of a backend and frontend. The former supports data analytics and service provisioning algorithms, while the latter offers configuration functionality to operators. Specially, in the backend, the historical traffic data from operator database is preprocessed to the proper time series format and then fed to the clustering and classification modules for extracting workload patterns. With the pattern database, AidOps predicts daily patterns using a pattern selection module. Finally the predicted time-varying patterns are used



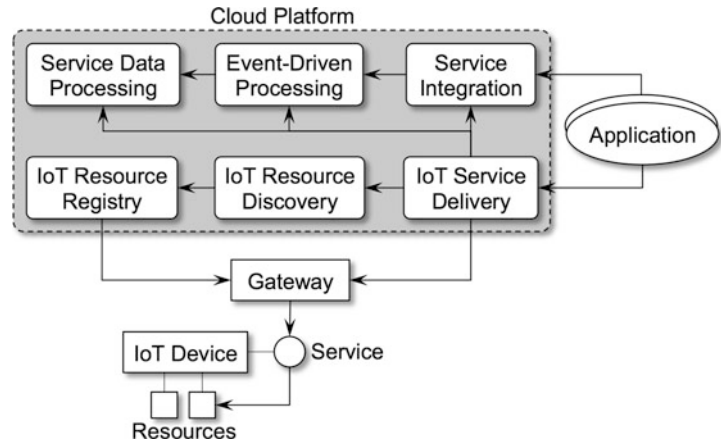
Data-Driven Service Provisioning, Fig. 2 A conceptual architecture of service-oriented decision support systems presented by Demirkan and Delen (2013)

Data-Driven Service Provisioning, Fig. 3

AidOps architecture proposed by Lugones et al. (2017)



Data-Driven Service Provisioning, Fig. 4 An overview of the IoT service model in cloud platforms, as presented by Taherkordi et al. (2018)



to optimize the configuration solution that is demanded by service orchestrator.

Facing the convergence of IoT and cloud, it is significant to facilitate the service provisioning and management in large-scale applications, particularly under the scenario of smart cities. To model and expose cloud-based IoT services in an efficient manner, Taherkordi et al. (2018) proposed CARIoT as shown in Fig. 4, which is a software framework for real-time provisioning of cloud-based IoT services. The main idea behind CARIoT is to structure the description of data-driven IoT services and the service-generated

data, both real-time and historical, in a hierarchical perspective corresponding to the context model of end-user application. With CARIoT, the end-user applications thus can access IoT services and their data in an efficient way.

Important Issues

The important issues under the data-driven service provisioning can be summarized as but not limited to the following topics.

- Efficient models and solutions for querying and processing big data.
- Big data-based service modeling, delivery, deployment, and evolution.
- Analytic services for big data.
- Knowledge discovery over massive datasets.
- Social sensing and social network services.
- Reasoning over uncertain data and unreliable services.
- Big data based service ranking and recommendation.

- Taherkordi A, Eliassen F, McDonald M, Horn G (2018) Context-driven and real-time provisioning of data-centric IoT services in the cloud. *ACM Trans Internet Technol (TOIT)* 1(1):1–20
- Xu Z, Liang W, Xu W, Jia M, Guo S (2015) Capacitated cloudlet placements in wireless metropolitan area networks. In: *IEEE 40th conference on local computer networks (LCN)*. IEEE, pp 570–578
- Zheng Z, Zhu J, Lyu MR (2013) Service-generated big data and big data-as-a-service: an overview. In: *2013 IEEE international congress on big data (BigData Congress)*. IEEE, pp 403–410

Cross-References

- [Data-Driven Edge Computing](#)
- [Data-Driven Mobile Social Networks](#)
- [Data-Driven Resource Allocation](#)
- [Data-Driven Smart City](#)
- [Data-Driven QoE Analysis](#)

References

- Chesbrough H, Spohrer J (2006) A research manifesto for services science. *Commun ACM* 49(7):35–40
- Demirkan H, Delen D (2013) Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. *Decis Support Syst* 55(1):412–421
- Huang H, Guo S (2017) Service provisioning update scheme for mobile application users in a cloudlet network. In: *IEEE international conference on communications (ICC)*. IEEE, pp 1–6
- Huang H, Guo S, Liang W, Li K, Ye B, Zhuang W (2016) Near-optimal routing protection for in-band software-defined heterogeneous networks. *IEEE J Sel Areas Commun* 34(11):2918–2934
- Huang H, Guo S, Wu J, Li J (2017a) Service chaining for hybrid network function. *IEEE Trans Cloud Comput*, <https://doi.org/10.1109/TCC.2017.2721401>
- Huang H, Li P, Guo S, Liang W, Wang K (2017b) Near-optimal deployment of service chains by exploiting correlations between network functions. *IEEE Trans Cloud Comput*, <https://doi.org/10.1109/TCC.2017.2780165>
- Huang H, Guo S, Wang K (2018) Envisioned wireless big data storage for low-earth-orbit satellite-based cloud. *IEEE Wirel Commun* 25(1):26–31
- Lugones D, Aroca JA, Jin Y, Sala A, Hilt V (2017) Aidops: a data-driven provisioning of high-availability services in cloud. In: *Proceedings of the 2017 symposium on cloud computing*. ACM, pp 466–478

Data-Driven Smart City

Deze Zeng

School of Computer Science, China University of Geosciences, Wuhan, China

Synonyms

[Data-Driven Digital City](#); [Data-Driven Intelligent City](#)

Definitions

Data-driven smart city is a cyber-physical system (CPS) that exploits the digital data acquired from a variety of sources to enable intelligent city administration in various domains. The data could be in the forms of government data, citizens' personal data, real-time city surveillance data, etc. By intelligently analyzing these data, it facilitates various city administration decisions such as road planning, transportation scheduling, power control, market regulation, and facility management to promote the quality of urban life from various aspects, e.g., safety, health, greenness, comfort, and efficiency. The rapid development of information and communication technology (ICT) provides a lot of tools to realize the vision of data-driven smart city, covering all the process cycles in data-driven smart city applications, i.e., data acquisition, data communication, data storage, data analytics, decision-making, and actuation. Based on these technologies, many

data-driven smart city applications have been proposed, implemented, and applied in different areas.

Historical Background

The concept of smart city was first formally proposed by IBM in 2008. Thereafter, smart city has been widely put into practice. In 2009, IBM and Dubuque initiated the project to cooperatively build the first smart city (Naphade et al. 2011). In 2009, the US government also initiated the smart grid project to empower the power delivery with ICT technologies (Chen et al. 2009). In the same year, China initiated the launch of project “Sensing China”, which intends to ubiquitously explore the power of IoT technologies. In 2010, the EU’s Horizon 2020 program treat smart city as one of the main development directions (Li et al. 2015). In 2011, Japan announced the “i-Japan” strategy, which intends to incorporate digital technologies into daily lives (Paskaleva 2011). Besides, along with the development of various ICT technologies, the smart city was incrementally re-defined by incorporating the newly emerging technologies. Without doubt that IoT is the first driving force of smart city and it is also the key to enable data-driven smart city by digitalizing the things in the city. Cloud computing then was incorporated to bear the data processing tasks (Parwekar 2011). Due to the long delay and unpredictable jitter of cloud computing, edge computing then was introduced to support delay-sensitive data-driven smart city applications (Gu et al. 2017). Besides the infrastructure innovations, Artificial Intelligence (AI) technologies was also introduced to empower the data processing capabilities (Wu et al. 2016) and future networking technologies was incorporated to support vast connections in smart cities (Tong et al. 2015).

Enabling Technologies

Data-driven smart city leverages various information technologies such as Internet of

Things (IoT), cloud computing, big data, and machine learning for urban planning, construction, monitoring, and management towards green, comfort, efficient, and sustainable city development. Data-driven smart city basically is a CPS that is data-centric. Data generation, transmission, storage, and processing are the key phases in data-driven smart city applications. Many related enabling technologies have been proposed and applied in these phases. For example, IoT technologies support the data generation from the physical world. Various wireless communication technologies and standards provide a variety means of data transmission means. Cloud computing provision an extremely large pool of computation and storage resources and edge computing alternatively provision the computation and storage resources in the citizen proximity. A diverse of artificial intelligence (AI) technologies are available for analyzing the data to provide various demanded city administration decisions.

Internet-of-Things

IoT interconnects various kinds of data generation devices, e.g., radio frequency identification (RFID), sensors, capable of identification, location, tracking, monitoring to realize the synergy between the physical world and the cyber world. As one of the key enabling technologies in data-driven smart city, IoT is used to acquire the desired information in every corner of the city. For example, we can use IoT to surveil the city environment; to monitor the structure health of the road, bridge, and buildings in the city; to manage the grid of the city; and to watch the water distribution system in the city (Zeng et al. 2016b). India and IBM cooperatively developed a real-time water quality management system called “Swatchpaani,” which is powered by IBM’s Watson IoT. Many IoT devices are injected in the water distribution system to continuously monitor water quality via measuring pH, temperature, and the presence of substances (Sandha et al. 2016). The IoT data are essential to build data-driven smart city applications and shall be efficiently acquired and processed.

Wireless Communications

Data-driven smart city applications heavily rely on the data, not only from the IoT devices but also from the human beings (e.g., social network information). The data are usually acquired via wireless communications, which have been developed for several decades. Many wireless communication standards targeted at different application fields have been proposed. Previously, wireless communication or mobile communication standards are usually designed with objectives like high throughput, low latency, high reliability, and low energy. To data-driven smart city applications, it is highly recommended that the new standards shall be with the capability of vast connections such that not only human-attached devices (e.g., mobile phone) but also a large number of IoT devices can be simultaneously connected in the 5G networks for seamless integration of the physical world and cyber world (Tong et al. 2015). The IoT business even may become the main driving force of the 5G networks, which shall support as the base infrastructure for data-driven smart city.

Cloud Computing

Cloud computing provisions computing and storage services in the datacenter in a centralized manner. There are three kinds of cloud computing service provision paradigms, i.e., software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS). To data-driven smart city application development, it is also desirable to host the data processing and storage functions in the cloud. Specially, many existing cloud services can be also utilized to mashup more powerful smart city applications. For example, we can mashup the map service with the traffic information acquired via various kinds of sensing devices (e.g., carmounted GPS, surveillance cameras, radar devices) to build an augmented city traffic map for a real-time understanding of the city traffic condition. Based on such augmented map, smart traffic control can be applied to manage the traffic lights for more efficient traffic control. To utilize the huge computing resources in the cloud datacenter, many different kinds of programming frameworks, e.g.,

MapReduce, Storm, are also available. These frameworks provide promising solutions for big city data processing to build data-driven smart city applications.

Edge Computing

Edge computing, as complementary to cloud computing, pushes forward the cloud computing to the network edge such that the computing can be within the proximity of end users for fast data processing (Zeng et al. 2016a). The edge computing servers can be colocated with the access networks, e.g., base stations. Edge computing is characterized by low latency and high bandwidth. Such features are consistent with the requirements of many data-driven smart city applications that require fast data processing and prompt response. For example, one of the most promising smart city applications is auto driving. A huge volume of data shall be processed in a prompt way and cloud computing is not desirable any more with the consideration of limited bandwidth, unpredictable long network delay, and unstable network connections. Resorting to edge computing to provision the computing resources at the network edge becomes a promising solution.

Artificial Intelligence

AI technologies have been developed for several decades and many different machine learning technologies with different capabilities have been proposed and applied in different domains, e.g., decision tree, support vector machine, neural network, Bayesian networks, reinforcement learning. Deep learning, as a newly emerging machine learning technology, further enhances the capability of machine learning in dealing with more complicated tasks. To data-driven smart city applications, AI plays a critical role in data-driven smart cities as these technologies have shown great potential in data analytics for prediction, inference, and decision-making. Specially, modern AI technologies are usually integrated with big data processing technologies so as to excavate the value of data to enable data-driven smart city (Wu et al. 2016). For example, we may apply deep reinforcement learning to

intelligently control the traffic lights according to the real-time information from the augmented map (Wu et al. 2017).

Application Domains

Having the advanced information, communication, and control technologies as backbones, data-driven smart city can be applied in a set of different application domains, including, but not limited to, smart transportation, smart healthcare, smart environment, smart living, and smart governance.

Smart Transportation

Smart transportation mainly focuses on the informationization of urban traffic. It provides data support for decision-making activities through the collection, transmission, and analysis of traffic related data, including both real-time and historical ones. Through the effective integration of various information technologies in traffic management, it provides the citizens with intelligent routine planning, traffic light control, road network structure suggestions. In the upcoming auto driving era, smart transportation shall further perform as the base infrastructure to support the intelligent operation, planning, and management of auto driving cars.

Smart Healthcare

Smart healthcare utilizes IoT technology to realize the interaction between the patients, medical institutions, and doctors to realize real-time, intelligent, and automated dynamic healthcare services (Gu et al. 2017). For example, smart healthcare centers can integrate patient health records from multiple sources to enrich the healthcare database for more advanced disease diagnosis. Smart healthcare also relies on data from connected medical devices, some of which are usually wearable sensors. These sensing data, combined with the surrounding environment information, are essential to many disease diagnosis and treatment. With smart healthcare, the citizens can receive more accurate treatment, thanks to the big healthcare data, and

in a more efficient and on-demand way, thanks to the tele-health system supported by the advanced information technologies.

Smart Environment

Smart environments aim to improve the sustainability of cities, the quality and safety of citizens' lives. Ubiquitous sensing and intelligent climate management are jointly applied in smart environment applications. This enables the timely detection of unhealthy or hazardous conditions and allows the authorities to react accordingly. For example, smart water distribution system utilizes contaminant detection sensors to detect water pollution or contamination and applies depollution actuation to ensure the safety of the citizens. Furthermore, sensor networks can detect natural disasters such as earthquakes, volcanic eruptions, tornadoes, floods, and forest fires. Early warning systems can significantly contribute to limiting casualties and reducing property damage.

Smart Living

In some areas, smart living offers intelligent management of various appliances and utilities to create comfortable homes and improve energy efficiency simultaneously (Liu et al. 2015). Smart living enables remote control of home appliances, especially intelligently in conjunction with the information about the climate change, energy provision, people's daily schedule, and social network data. In the community (or building), smart living applications also intelligently manage waste recycling, social networking, and parking to provide a smart community (or building) with comfortable lives.

Smart Governance

Smart governance aims to build a new and innovative government service platform to achieve a harmonious society with government guidance, market, and citizen participation, making public services more refined, socialized, and intelligent. For example, open data enables citizens, third-party developers, and the city council to access and cross-reference different data sources, which can increase both transparency and efficiency. Smart governance services allow citizens to

complete most interactions automatically, e.g., booking weddings, applying for social housing, resident parking permits, or schools. It also allows citizens to get involved in the city planning and development processes to improve the quality of city life.

Cross-References

- [Big Data Networks](#)
- [Big Network Data](#)
- [Data-Driven Edge Computing](#)
- [Data-Driven Mobile Social Networks](#)
- [Data-Driven Pattern Prediction](#)
- [Data-Driven Resource Allocation](#)
- [Data-Driven Security](#)
- [Data-Driven Service Provisioning](#)

References

- Chen S, Song S, Li L, and Shen J (2009) Survey on smart grid technology. *Power system technology* 8:1–7.
- Gu L, Zeng D, Guo S, Barnawi A, Xiang Y (2017) Cost efficient resource management in fog computing supported medical cyber-physical system. *IEEE Trans Emerg Top Comput* 5(1):108–119. <https://doi.org/10.1109/TETC.2015.2508382>
- Li Y, Lin Y, Geertman S (2015) The development of smart cities in China. In 2015 14th International Conference on Computers in Urban Planning and Urban Management, Cambridge, Massachusetts USA, pp 7–10.
- Liu L, Liu Y, Wang L, Zomaya A, Hu S (2015) Economical and balanced energy usage in the smart home infrastructure: a tutorial and new results. *IEEE Trans Emerg Top Comput* 3(4):556–570. <https://doi.org/10.1109/TETC.2015.2484839>
- Naphade M, Guruduth B, Colin H, Jurij P, and Robert M (2011) Smarter cities and their innovation challenges. *Computer* 44(6):32–39. <https://doi.org/10.1109/MC.2011.187>
- Parwekar, P (2011) From internet of things towards cloud of things. In 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011), IEEE, pp. 329–333.
- Paskaleva, KA (2011) The smart city: A nexus for open innovation? *Intelligent Buildings International* 3(3):153–171.
- Sandha SS, Randhawa S, Srivastava B (2016) Blue water: a common platform to put water quality data in India to productive use by integrating historical and real-time sensing data. Technical report, IBM Research Report. IBM
- Tong W, Ma J, Huawei PZ (2015) Enabling technologies for 5g air-interface with emphasis on spectral efficiency in the presence of very large number of links. In: 2015 21st Asia-Pacific conference on communications (APCC), Kyoto, pp 184–187. <https://doi.org/10.1109/APCC.2015.7412508>
- Wu J, Guo S, Li J, Zeng D (2016) Big data meet green challenges: big data toward green applications. *IEEE Syst J* 10(3):888–900. <https://doi.org/10.1109/JSYST.2016.2550530>
- Wu C, Parvate K, Kheterpal N, Dickstein L, Mehta A, Vinitsky E, Bayen AM (2017) Framework for control and deep reinforcement learning in traffic. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC), Yokohama, pp 1–8. <https://doi.org/10.1109/ITSC.2017.8317694>
- Zeng D, Gu L, Guo S, Cheng Z, Yu S (2016a) Joint optimization of task scheduling and image placement in fog computing supported software-defined embedded system. *IEEE Trans Comput* 65(12):3702–3712. <https://doi.org/10.1109/TC.2016.2536019>
- Zeng D, Gu L, Lian L, Guo S, Yao H, Hu J (2016b) On cost-efficient sensor placement for contaminant detection in water distribution systems. *IEEE Trans Ind Inf* 12(6):2177–2185. <https://doi.org/10.1109/TII.2016.2569413>

Data-Oriented QoE Management

- [Data-Driven QoE Measurement](#)

Data-Oriented Wireless Sensor Networks

- [Information-Centric Wireless Sensor Networks](#)

DCN

- [Data Center Network in Cloud Computing](#)

DCN Traffic Scheduling

- [Load Balancing in Data Center Networks](#)

DDoS

- [New Variants of Mirai and Analysis](#)

Decision-Making

- [Neighborhood-Based ICT-Driven Support \(NIS\) System of Emergency Medical Services \(EMS\) for Elderly People in Hyper-Aged Societies](#)

Dedicated Short-Range Communication (DSRC)

- [Wireless Access in Vehicular Environment](#)

Definition of Wireless Sensor Network

- [Wireless Sensor Network Security](#)

Delay Tolerant Networks

- [Data Dissemination in Delay Tolerant Networks with Geographic Information](#)

Delay-Tolerant Network Routing

Yue Cao
School of Computing and Communications,
Lancaster University, Lancaster, UK

Synonyms

[Mobility](#); [Opportunistic encounter](#); [Routing](#)

Definition

Delay-tolerant network routing utilizes “store-carry-forward” mechanism to relay data, through asynchronized and opportunistic encounters between pairwise nodes. The criterion on selection of relay node is based on various strategies with network topology or geographic information.

Historical Background

Due to the characteristic of challenged environment suffering from frequent link disruption, sparse network density, and limited device capability, routing algorithms designed for mobile ad hoc networks (MANETs) cannot perform effectively under these constraints. This is because the availability of contemporaneous end-to-end connectivity is essential for conventional routing algorithms such as ad hoc on-demand distance vector (AODV) (Perkins et al. 2003) or dynamic source routing (DSR) (Johnson and Maltz 1996). However, such does not prevent bridging communication between the disconnected areas by following the technologies from delay-tolerant networks (DTNs) (Cao and Sun 2013) to overcome these difficulties using the store-carry-forward (SCF) routing mechanism.

In DTNs, mobile nodes can communicate with each other, even if the contemporaneous end-to-end connectivity is unavailable. Furthermore, the global knowledge about network is not mandatory for network nodes in DTNs. The lack of contemporaneous end-to-end connectivity prevents the conventional routing algorithms designed for MANETs from working effectively. Here, the Bundle Protocol (Scott and Burleigh 2007) borrowing from the concept of Email protocol is proposed by the Internet Research Task Force (IRTF) Delay-Tolerant Networking Research Group (DTNRG) to behave as a convergence layer protocol on top of the Transmission Control Protocol (TCP) layer to enhance the transmission reliability in DTNs.

Delay-Tolerant Network Routing, Table 1 Comparisons between routing in MANETs and DTNs

	Routing in MANETs	Routing in DTNs
End-to-end connectivity	Contemporaneous	Frequently disconnected
Delivery delay	Short	Long
Transmission reliability	High	Low
Routing behavior	Symmetric	Asymmetric

It is worth to note that other terms, e.g., intermittently connected networks (ICNs) (Khabbaz et al. 2012) or opportunistic networks (Pelusi et al. 2006), were proposed by other articles, while they are normally exchangeable with DTNs. As illustrated in Table 1, routing in DTNs suffers more from long delivery delay than that of MANETs due to the asymmetrical routing behavior and low transmission reliability while taking into account the limited encounter duration due to the lack of contemporaneous end-to-end connectivity.

Foundations

Useful Terms

Encounter opportunity: It is an encounter between pairwise nodes. Specifically, an encounter opportunity is regarded as a tuple consisting of (d, m, b) , where d is the time duration of an encounter, m is a set of messages requested for transmission, and b is the bandwidth speed of DTN device. For reliable transmission, the total size of messages m being transmitted during an encounter opportunity should not exceed the maximum volume of the encounter opportunity, which is determined by $d \times b$.

Store-carry-forward: When a node carries a message, while there is no contemporaneous end-to-end path to its destination or even a connectivity to any other node, this message would be stored locally. The sender would wait for the upcoming encounter opportunity with other nodes for message relaying.

Candidate node: Based on the definition of store-carry-forward, the encountered node selected as the message relay is defined as the candidate node for this message.

Inherent Challenges

Bandwidth: This factor determines the number of messages that can be transmitted at each encounter opportunity. For instance, if the bandwidth of a DTN device is sufficient to transmit all the requested traffic load within a given encounter duration, then this is reasonable. However, if the traffic load increases due to a large number of users or a larger size of messages being transmitted, then the unsuccessful transmissions due to insufficient encounter duration should be taken into account. Therefore, to estimate the number of messages that can be successfully transmitted is useful to reduce the number of aborted messages due to insufficient encounter duration. In addition, to transmit messages according to a corresponding priority is beneficial to utilize the bandwidth.

Buffer space: The sufficient buffer space is essential for the carried messages, since they would be buffered for a long period time until the upcoming encounter opportunity is available. In light of this, to discard the least important message due to buffer space exhaustion is beneficial to utilize the buffer space.

Energy: A DTN device often has limited energy and cannot be connected to the power supplier easily. Energy is required for transmitting, receiving, and storing messages and performing routing process. Hence, the DTN routing algorithms which transmit few messages and perform less computation are more energy efficient.

Mobility Model

Apart from the inherent challenges, mobility factor (Grossglauser and Tse 2002) describes the variation of movement and plays an important role in routing performance in DTN. Therefore, it is desirable to emulate the movement pattern of

the targeted real-world applications in an appropriate way. Otherwise, the conclusions and observations drawn from the results may be misleading. In light of this, it is necessary to select the appropriate underlying mobility model while evaluating the routing performance. For example, the mobile nodes under the random Waypoint (RWP) mobility model would behave differently from the group-based mobility model. In reality, since it is difficult to obtain the global knowledge about the distribution of encounter probability or inter-meeting time, knowledge and assumption regarding mobility model are more crucial.

Evaluation Metrics and Routing Objective

Delivery ratio: It is given by the ratio between the number of delivered messages and the number of generated messages.

Overhead ratio: It is given by the ratio between the number of message transmissions required for delivery and the total number of messages delivered.

Delivery delay: It is given by the time duration between the message generation and their delivery.

The routing objective provides a trade-off between maximizing the delivery ratio and minimizing the overhead ratio. On one hand, the ideal case of delivering the message before its given lifetime with the lowest overhead ratio is to keep this message until the destination is in proximity. While on the other hand, the effective approach to maximize the message delivery ratio is to relay this message at each encounter opportunity (if necessary, taking into account the candidate node selection strategy). Although it is expected that the applications of DTNs are inherently tolerant to the long delivery delay, this does not mean they would not benefit from short delivery delay; thus this should be a still target with the given message lifetime.

Summary of Routing Protocols in DTN

Literature works have formed the following two mechanisms during routing procedure:

Single copy-based: The data is normally relayed through several candidate nodes toward

the receiver. During the routing procedure, there is no message copy generated. The performance of routing protocols under this mechanism is substantially influenced by nodal mobility model, where infrequent encounter opportunity between nodes may result in failure of data delivery.

Multiple copy-based: A number of copies of the message will be generated. This is driven by the aim that at least one of these copies will be delivered, through multiple paths formed by opportunistically selected candidate nodes. The routing performance with this mechanism is greatly improved particularly in networks which are quite sparse. The downside is a significantly increased routing overhead incurred by exchanging of message copies.

Naturally, based on what knowledge used for making routing decision, DTN routing protocols are classified into following two branches.

Topology-Based Routing Protocols in DTNs

There have been many previous works address this technique branch, in (Cao and Sun 2013) several metrics have been reviewed to guide with design of routing protocols.

The one-hop encounter featured routing protocols rely on one-hop prediction, e.g., direct encounter probability or encounter interval/frequency, to qualify the candidate node, while the time-varied shortest path-based routing protocols expect a global view about the knowledge such as queue size and inter-meeting time of other nodes in the network. Using these information, a classic Dijkstra's approach forms a path consisting of opportunistic encounters with the time-varying property. Apart from this, the social relationship-based metric considers the social tie among the linked nodes. As such, many existing methods from online social networks have been applied in DTN routing.

Besides, the assumption of traditional congestion control approach is based on the contemporaneous end-to-end connectivity, enabling both the congestion feedback and control information to be received timely and successfully. Since it is difficult to perform this end-to-end-based approach due to the constraints in DTNs, the hop-by-hop approach is appropriate instead.

This gears the routing protocol with the nodal resources constraints like buffer and bandwidth to discard incoming/carried data, reduce the number of copies of data, and alternatively transfer the data to another temporary candidate node to help hold the data.

In summary, the above-featured routing protocols estimate how reliable that pairwise nodes will encounter and communicate. A good candidate reflects it will meet the current data carrier with a short time interval but high frequency, of course a long communication duration in each encounter is a plus for reliable data transmission.

Geographic-Based Routing Protocols in DTNs

Geographic routing (Karp and Kung 2000), also called position-based routing, requires that each node can determine its own location and the source is aware of the location of destination. Different from topological routing, geographic routing exploits the geographic information instead of topological connectivity information for message relay to gradually approach and eventually reach the intended destination. However, the design of geographic routing (Wang et al. 2016) in DTNs inevitably faces the following challenges:

- In DTNs, the network node which is currently closer to the destination may not be still a good candidate node in the future. This is because that it may move away from the destination within a short time, particularly concerning the high mobility in sparse network.
- Concerning the mobility of destination, it limits the feasibility of applying a centralized location server to track the real-time location of mobile destination in sparse networks. This happens due to the fact that there is a long delay to request/reply the information from location server in DTNs, while the obtained location information may be outdated and inaccurate for making routing decision.
- Conventional geographic routing protocols in MANETs rely on the high network density, which is infeasible in DTNs because there are

insufficient number of encountered nodes for handling this problem. The local maximum problem generally refers to the case that there is no other candidate nodes in proximity, as such the message will be kept locally. In the worst case, the data will expire and get discarded locally.

Key Applications

The space application of DTNs is for InterPlanetary Networks (IPNs) (Caini et al. 2011) with a low network dynamic. In mobile wireless networks, the terrestrial applications of DTNs have been envisioned for vehicular ad hoc networks (VANETs) (Pereira et al. 2012), airborne networks (Jonson et al. 2008), packet-switched networks (PSNs), and underwater networks (UWNs) (Wei et al. 2014).

IPNs: Here, the high bit error rate and long delay of wireless communication condition in deep space scenario are of major challenges.

VANETs: Facing frequent link disruption between vehicles, the exchange of intervehicle information (road traffic, motion, etc.) can achieve safer and efficient driving in VANETs.

ANs: Apart from the major applications in the military work previously, unmanned aerial vehicles can also achieve the image acquisition with high imaging resolution, incorporate with satellite sensing systems to monitor environments in disaster scenario, and also monitor in the development of smart cities.

PSNs: Concerning human routine, with continued and dramatically increased interests in the civil applications alongside collaboration with adjacent sectors, PSNs make use of both human mobility and social relationship connectivity to relay the message between mobile users.

UWNs: Acoustic communication systems differ from terrestrial telemetry, due to differences in system geometry and environmental conditions. With the autonomous underwater vehicles delivering data to sink node, UWNs are envisioned for oceanographic applications such as pollution monitoring, etc.

Future Directions

It has been years since DTN routing has been proposed with numerous contributions investigated. Right now, given the emerging trend of 5G technology, DTN routing is able to support the network performance as complementary role when cellular network infrastructures are in failure. Artificial intelligence (AI) algorithms should be adequately investigated to optimize the routing decision, e.g., machine learning and Bayesian and reinforcement solutions are worth the investigation. Particularly, these technologies would help to predict network condition and variation, so as to switch toward DTN routing or long-range ubiquitous communication. In addition to unicasting routing protocols being mainly studied, multicasting and anycasting should be promoted as well (Cao et al. 2018), e.g., implementing an anycast-based routing protocol in VANETs.

Cross-References

- [Data Dissemination in Delay Tolerant Networks with Geographic Information](#)
- [Routing for Vehicular Ad Hoc Networks](#)
- [Vehicle Ad-Hoc Networks: Services, Security, and Privacy](#)

References

- Caini C, Cruickshank H, Farrell S, Marchese M (2011) Delay-and disruption-tolerant networking (DTN): an alternative solution for future satellite networking applications. *Proc IEEE* 99(11):1980–1997
- Cao Y, Sun Z (2013) Routing in delay/disruption tolerant networks: a taxonomy, survey and challenges. *IEEE Commun Surv Tutorials* 15(2):654–677
- Cao Y, Wang T, Zhang X, Kaiwartya O, Eiza MH, Putrus G (2018) Towards anycasting-driven reservation system for battery switch based electric vehicle charging. *IEEE Syst J*
- Grossglauser M, Tse DN (2002) Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Trans Netw* 10(4):477–486
- Janardanan Kartha J, Jacob L (2015) Delay and lifetime performance of underwater wireless sensor networks with mobile element based data collection. *Int J Distrib Sens Netw* 11(5):128757
- Johnson DB, Maltz DA (1996) Dynamic source routing in ad hoc wireless networks. In: *Mobile computing*. Springer, Boston, pp 153–181
- Jonson T, Pezeshki J, Chao V, Smith K, Fazio J (2008, November) Application of delay tolerant networking (DTN) in airborne networks. In: *Military communications conference, 2008. MILCOM 2008. IEEE, Piscataway*, pp 1–7
- Karp B, Kung HT (2000, August) GPSR: Greedy perimeter stateless routing for wireless networks. In: *Proceedings of the 6th annual international conference on Mobile computing and networking*. ACM, New York, pp 243–254
- Khabbaz MJ, Assi CM, Fawaz WF (2012) Disruption-tolerant networking: A comprehensive survey on recent developments and persisting challenges. *IEEE Commun Surv Tutorials* 14(2):607–640
- Pelusi L, Passarella A, Conti M (2006) Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. *IEEE Commun Mag* 44(11):134–141
- Pereira PR, Casaca A, Rodrigues JJ, Soares VN, Triay J, Cervello-Pastor C (2012) From delay-tolerant networks to vehicular delay-tolerant networks. *IEEE Commun Surv Tutorials* 14(4):1166–1182
- Perkins C, Belding-Royer E, Das S (2003) Ad hoc on-demand distance vector (AODV) routing (No. RFC 3561)
- Scott K, Burleigh S (2007) Bundle protocol specification (No. RFC 5050)
- Wang T, Cao Y, Zhou Y, Li P (2016) A survey on geographic routing protocols in delay/disruption tolerant networks. *Int J Distrib Sens Netw* 12(2):3174670
- Wei K, Liang X, Xu K (2014) A survey of social-aware routing protocols in delay tolerant networks: applications, taxonomy and design-related issues. *IEEE Commun Surv Tutorials* 16(1):556–578

Demand-Side Economies of Scale

- [Network Effects](#)

Densification of Base Stations

- [Densification of Wireless Networks, from Few to Massive](#)

Densification of Wireless Networks, from Few to Massive

Shuyi Chen
Harbin Institute of Technology, Harbin, China

Synonyms

Densification of base stations

Definitions

Network densification is the process of deploying more cells during the development of wireless communication industry, which is by far the greatest contributor to the evolution of network capacity.

Historical Background

The 1970s witnessed the emergence of the first-generation cellular network. Since then, wireless communications have developed into one of the largest sectors of the telecommunications industry and improved humans' living experience in nearly every aspect. It is worth noting that there were only 10 million subscribers in 1990, mostly using analog FM technology (Rappaport et al. 2002). In 2017, the 5 billion mobile subscriber milestone was reached, and a further 620 million subscribers will be added by 2020, reaching almost three quarters of the global population (GSMA Intelligence 2017). With the realization of the Internet of Things (IoT), smartphones are not the only connected devices. Now, the machine-to-machine market is undergoing a fast transformation over the wireless cellular networks. In 2017, IoT connections have outnumbered traditional smartphone connections.

Meanwhile, since the first car phone call was made in 1979, multiple services have been supported by the cellular networks apart from the voice services. For example, the first text message

was sent in December 3, 1992, with a greeting of Merry Christmas through short message service (SMS) of the second generation of mobile technology (2G). Later, multimedia message service (MMS) standard extended the core SMS capability, allowing messages with multimedia content. Mobile Internet data requests were handled by Wireless Application Protocol (WAP) or i-mode in the 1990s, and the first WAP-enabled service hit Europe in October 1999 and i-mode was launched in Japan in February 1999. At that time, checking e-mail or daily news was the premiere application. Now, nearly three quarters of smartphone users globally watch online videos on their phones. A total of 5 exabytes video traffic was transmitted monthly in 2016, and the number was estimated to reach 50 exabytes by 2022 (GSMA Intelligence 2017). Accordingly, the theoretical transmission data rate has increased from 9.6 kbit/s for Global System for Mobile Communications (GSM) in 2G to 1 Gbit/s for Long-Term Evolution (LTE)-Advanced in 4G.

Even though only four decades have passed since the emergence of the first-generation cellular network, wireless communication system has experienced tremendous changes to support such large traffic volume from a multitude of users with diverse service requirements. One obvious change is the development of mobile phones, which originated as car phones, turned to huge heavy portable devices, and finally became modern smartphones.

The evolution of wireless infrastructure equipment, compared to mobile phones, is not that closely linked with humans' life. However, wireless infrastructure equipments have experienced a dramatic change not only in the structure but also in quantitative terms. Looking over the development of wireless network, introducing more cells is by far the greatest contributor to the evolution of network capacity. From 1960 to 2010, deploying more cells has increased the capacity by 2000 times, while using extra spectrum and improving spectral efficiency totally only contribute 50 times in capacity increases (Agilent Technologies 2013).

The intent of this entry is to explore the role of base station densification during the evolu-

tion of wireless system. Initially, increasing the density of macro cells is to meet the coverage and capacity requirement in earlier cellular network. Later, the concept of small cell has been proposed to meet high capacity requirement in certain areas. Since then, the number of small cells has increased dramatically, and this trend is forecasted to continue as the ultradense small cell network has become essential for 5G network. Focusing on the number of base stations, the rest of this entry discusses the benefits and challenges of network densification in homogeneous and heterogeneous networks, respectively.

Homogeneous Networks

In the early days of cellular network, only macro cells were deployed, which formed a homogeneous network. The typical coverage of a macro cell ranges from 1 to 35 km. In an initial network, planners typically aim at using fewest possible base stations to provide the maximum coverage. As subscribers increase, planners need to optimize network capacity, and cell size should be as small as possible to accommodate additional capacity. Thus, the development of homogeneous network can be roughly divided into two stages based on the different planning focuses, i.e., coverage planning and capacity planning.

Stage 1: Coverage Planning

The initial period of the first-generation cellular network can clearly show the premier coverage concern of network planners. For example, 88 cell sites were deployed in the first commercial cellular network launched by Nippon Telegraph and Telephone (NTT), a Japanese telecommunications company, on December 1, 1979, with a fully functional network covering the 23 wards of Tokyo. Fifty-six cells and 36 cells were distributed, respectively, by AT&T and Motorola to cover New York city using advanced mobile phone system (AMPS). However, in the 1970s and 1980s, even though the rational coverage was claimed in different countries, most mobile phones only worked when they were close to a base station.

The first generation of analog cellular network was gradually replaced by the second generation of digital network since the commercial launch of the GSM service in 1991 with limited coverage of larger cities/airports in Europe, followed by a complete linking of areas by 1995. Since Telstra Australia became the first non-European operator to sign the GSM in 1993, GSM networks went live globally, and the milestone of 1 billion subscribers was reached in 2004. Even though 2G services are gradually replaced by new technologies, for example, NTT DoCoMo in Japan terminated its 2G services on March 31, 2012, 2G network still shares the highest coverage rate and the largest connected population, which is 40% of total connections in 2017 (GSMA Intelligence 2017).

Only four decades have passed since the emergence of 1G, and mobile network has almost reached global coverage. Today, only a quarter of the global population lack access to wireless network, and the majority of these uncovered population live in the rural regions of Asia and sub-Saharan Africa, and it is uneconomical to extend coverage of these area due to the high fixed costs of laying network infrastructure to serve sparsely distributed population with low purchasing power.

Stage 2: Capacity Planning

With the increasing number of subscribers, only ensuring the network coverage cannot meet the transmission requirements from massive users. For example, theoretically, 30 cells are needed in GSM system to support 10,000 subscribers; if the grade of service (GOS) is 2%, and the number of available channels is 24 (Ericsson 1998). When the number of subscribers surpasses the system limit, network planners need to make amendments to the existing system to increase the network capacity.

There are multiple approaches to improve the network capacity, and one direct method is to deploy more macro cells. However, it is too costly to rebuilt a new network with more base stations. In this case, applying cell splitting becomes a practical approach. Cell splitting is the process of subdividing a congested cell into smaller

cells such that each smaller cell has its own base station with reduced antenna height and reduced transmitter power. This method increases the capacity of a cellular system and minimizes the need to modify the existing cell parameters. However, there are challenges in increasing the capacity by cell splitting. As more cells are deployed, handoffs are more frequent, and the process of channel assignment becomes more difficult. Meanwhile, as all cells are not split simultaneously, special care should be taken for proper resource allocation.

Apart from splitting macro cells, adding sectors is the other way to increase network capacity, which replaces the omnidirectional antenna at each base station by three (or six) sector antennas of 120 (or 60) degrees opening (Babich and Vatta 2002). Correspondingly, the channel set serving this cell has also been divided, and thus each sector is allocated one-third (or one-sixth) of the available number of channels. Cell sectorization is less expensive than cell splitting, as no new base stations are required. Comparing to using omnidirectional antenna, this technique can reduce co-channel interference and thus increase the capacity. For example, comparing to using omnidirectional antenna, the network capacity is 1.5 times by using 60-degree sectored antennas (Small Cell Forum 2016). Meanwhile, using cell sectorization enables the reduction of the cluster size and provides an additional freedom in assigning channels. However, by dividing the bigger pool of channels into smaller groups, trunking efficiency is decreased due to sectoring, and the handoffs among different sectors are essential.

Heterogeneous Networks

Apart from adding more sectors or macro cells, an alternative approach to expand an existing macro-network is to introduce different kinds of small cells, such as femtocells, picocells, and microcells. Small cells are low-powered radio access nodes with a range of 10 meters to a few kilometers. The deployment of small cells within the coverage of macro cells can improve the

area spectrum efficiency greatly by reusing the same frequencies many times within a geographical area. The result is a heterogeneous network (HetNet) with large macro cells guaranteeing the network coverage and small cells providing increased capacity per unit area. The concept of small cell has already been proposed in GSM and supported by the LTE specifications since the beginning of LTE. Since then, HetNet has replaced the position of homogeneous network gradually and become the basic network structure for modern wireless system. The development of HetNet can be roughly divided into two stages based on the density of small cells, which will be discussed, respectively.

Stage 1: Sparse Small Cells

The primary purpose of adding small cells is to increase capacity in hot spots with high user demand or to fill the coverage gap in bland spots and cell-edge areas, both indoors and outdoors. Thus, during the early period of HetNet, small cells are deployed sparsely and separately.

The main challenge of deploying small cells within the coverage of macro cells is to tackle the interferences between macro and small cells, which are also referred as cross-tier interferences. A number of approaches can be used to mitigate the cross-tier interferences, and the three dominant methods are inter-cell interference coordination (ICIC), carrier aggregation (CA), and coordinated multi-points (CoMP), which have all been added into 3rd Generation Partnership Project (3GPP) LTE specifications. ICIC and its successors, enhanced ICIC (eICIC), and Further enhanced ICIC (FeICIC) were introduced in 3GPP Release 8, 3GPP Release 10, and 3GPP Release 11, respectively. The former concentrates on the frequency domain to decrease the interference by lowering the power of a part of subchannels of macro cells, and the latter two introduce a blanking of subframes of the macro cells in the time domain to mitigate the interferences. Apart from ICIC-related approaches, the other method is to apply CA with cross-carrier scheduling, which enables that user equipments can connect

to different base stations (macro cell and small cell base stations) on different carriers on the physical downlink control channel (PDCCH). CoMP, on the other hand, focuses on the collaboration between macro cells and small cells. For example, data can be transmitted simultaneously from different base stations to the same user equipment in the same resource block.

At the end of the previous section, the main challenge of achieving global coverage has been pointed out, which is the uneconomical factor to build network infrastructures in rural or remote areas to serve sparsely distributed populations with low purchasing power. In this case, the usage of small cell technology represents an economical alternative. For example, SoftBank Mobile, the Japanese mobile operator, has installed thousands of small cells throughout rural Japan by 2016. In the UK, Vodafone's rural Open Sure Signal project also deploys small cell in rural areas, and each unit provides up to 500 m of 3G coverage.

Stage 2: Dense Small Cells

Initially, few small cells are deployed sparsely and separately to provide high capacity in hot spots or fill the coverage gap. But now a large number of small cells are distributed where demand dictates, such as in cities, remote areas, workplaces, and leisure locations, and are starting to appear in vehicles and even drones. At the end of 2012, the number of small cells has surpassed 6 million (6,069,224), which outnumbered that of macro cells (5,925,974). Meanwhile, the density of small cell increases with time, and the hyperdense small cell network is believed to be essential to address the data surge in future network. In 2014, among the 207,000 urban small cells, about half of them (118,000) were sparsely distributed, and there is no hyperdense small cell network (Rethink Technology Research 2017). In 2016, the number of urban small cells reached 1 million, with one forth (256,000) sparse small cells and one tenth (112) hyperdense small cells (Rethink Technology Research 2017). With the start of rollout of 5G hyperdense zones of

capacity, the installed small cells will leap to about 14.5 million by 2022 (Small Cell Forum 2016), with 2 million hyperdense small cells (Rethink Technology Research 2017).

There are multiple challenges with the densification of small cells, and the construction of small cell backhaul network consistently appears high on the list of operator perceived barriers (Small Cell Forum 2017). Both wired and wireless backhaul technologies can be applied to connect small cells and macro cells with a stable, affordable, scalable, and high data rate link. Wired backhaul solutions, such as fiber and copper, provide optimal capacity and reliable connections. However, the biggest restriction is that it depends on a pre-established infrastructure, and such infrastructure exists mostly in residential and enterprise area, that is to say, indoor scenario. Meanwhile, using public Internet connections presents a security risk. In area without pre-established infrastructure, operators have to build their own wired backhaul network, which is time-consuming and costly. For example, it takes about 2 years for the Verizon Communications, an American telecommunication company, to establish the small cell network in several large cities in the USA, as every single cell has fiber backhaul to a macro cell. In cases like these, operators in need of quick deployments will generally choose wireless solutions, such as traditional microwave (6–42 GHz), sub 6 GHz microwave, unlicensed millimeter wave (60 GHz), and E-band microwave (70–80 GHz). Even though all these solutions for small cell deployments are available on the market, there is no uniform standard for this fast-growing market.

Conclusion

Only four decades have passed since the emergence of the first analog cellular network, and now there are nearly 7 million macro cells and more small cells globally. The densification of base stations is believed to be the greatest contrib-

utor to the evolution of network capacity. In the early period of cellular network, only macro cells existed and constituted a homogeneous network. One way to expand an existing macro-network, while maintaining it as a homogeneous network, is to densify the network by adding more sectors per cell (cell sectorization) or deploying more macro cells (cell splitting). However, reducing the site-to-site distance in the macro-network can only be pursued to a certain extent, and an alternative is to construct a heterogeneous network by adding different kinds of small cells. The main challenge in this stage is to tackle the interference between small cells and macro cells, and different approaches, such as ICIC, CA, and CoMP, have been added in LTE and LTE-A specifications to support heterogeneous network. With the fast-growing demand for high data rate, deploying high-density small cell network becomes an attractive solution, where the density of small cell base stations can be equal to or surpass the density of user equipments. There are many remaining challenges to deploy a high-density small cell network, and the construction of backhaul network consistently appears high on the list of operator perceived barriers. Different wired and wireless backhaul products are available in the market, and there is no uniform standard yet.

Even though every generation of cellular network has undergone the process of densifying base stations to meet the coverage requirement and then the capacity requirement, different approaches have been applied. 1G was purely consisted of homogeneous network, and the approaches of cell sectorization and cell splitting were proposed in 2G to increase the network capacity. Meanwhile, heterogeneous network planning was already used in 2G, where macro and small cells were allocated different frequencies to avoid the cross-layer interference. More effective solutions have been proposed to support heterogeneous network in 3G and 4G, and gradually, heterogeneous network has become the basic network structure. Since the emergence of small cell, its application scenarios have been enlarged continually, and deploying dense small cells has become a trend for 5G.

Key Applications

Network densification is essential in every generation of wireless communication system.

Cross-References

- [Cellular Networks: An evolution from 1G to 4G](#)
- [Deployment Strategies of Cellular Networks](#)
- [Heterogeneous Small Cell Networks](#)

References

- Agilent Technologies (2013) From angst to ambivalence to acceptance: the story of how cellular has evolved to embrace Wifi. Technical report
- Babich F, Vatta F (2002) Effects of sectorization on cellular radio systems: capacity with different traffic loads. *Wirel Pers Commun* 21:269–288
- Ericsson (1998) GSM system survey. Technical report
- GSMA Intelligence (2017) Global mobile trends 2017. Technical report
- Rappaport TS, Annamalai A, Buehrer RM, Tranter WH (2002) Wireless communications: past events and a future perspective. *IEEE Commun Mag* 40:148–161
- Rethink Technology Research (2017) Small cells service module 1: deployments and key trends. Technical report
- Small Cell Forum (2016) Capacity planning for HetNets. Technical report
- Small Cell Forum (2017) Network densification in the 5G era. Technical report

Deployment Strategies of Cellular Networks

Wei-Che Chien and Chin-Feng Lai
Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan, Republic of China

Synonyms

[Placement strategies of cellular networks](#)

Definitions

The deployment strategy of cellular network is a complete plan that includes evaluation mechanisms, processes, and algorithms to deploy base stations of cellular network. The aim of evaluation mechanisms is to assess the performance of deployment solution that the user can notice the improvements barely. The common definition of process is a series of actions to achieve deployment goals, and the algorithm is brought up for mathematical instructions or rules that are designed according to deployment goal.

Historical Background

In order to cover the communication range of large areas, the early mobile radio systems generally adopt high-power transmitters and then built them in higher geographical locations. However, the used frequency bands in the covered area could not be reused since they may cause interference when other transmitters used the same frequency band at the same time. In the healthcare field, some reports show the high-power waves may cause harm to human bodies. Therefore, the concept of a cellular network has been proposed to solve the above problem (Jump and Kirtane 1974). Each cell of a cellular network has a radio base station (RBS), which is called Node B in 3G and eNB in LTE. Radio base station will be connected to the core network through a fixed network (optical fiber) to replace the traditional single high-power base station. Because the coverage area of the cell is small, the transmitter does not need high power to transmit data so that the damage on human health also can be reduced. Different cells can use different bands to improve spectrum utilization. Because the distances between cells are far enough, the interference problems also can be avoided.

With the rapid development of the Internet, people's demand for the network has gradually increased. In order to achieve the goal that the people can access network anytime and anywhere, the network coverage rate has become

a problem that cannot be ignored. Therefore, many cells with different coverage areas have been developed. The coverage range of cells from large to small can be divided into macrocell, microcell, picocell, and femtocell. Among them, microcell, picocell, and femtocell are also called small cell. Macrocells must be operated under licensed spectrum. Small cells, which can be used to supplement for the noncoverage area of macrocells, can be operated in licensed and unlicensed spectrum. The deployment cost and restriction of services number are different with different cells, and the deployment position of RBS is limited by the environment. Furthermore, the spectrum allocation is also a challenge because interference will happen when the same band is used by two cells. Therefore, the deployment problem of cellular network is difficult and complex.

The deployment problems of cellular networks need to be solved by appropriate strategies according to the different goals, the demands, and the environment change. In order to provide a flexible and efficient deployment strategy, it is necessary to include evaluation mechanisms, processes, and algorithms. The evaluation mechanisms need be adjusted according to the deployment goals; for instance, the signal-to-interference-plus-noise ratio (SINR) is usually used to evaluate signal quality; Watts/kbps and Watts/user are used as metrics to quantify the performance of total network power consumption and the ratios (Koutitas et al. 2012). If multi-objective evaluation mechanisms have been adopted, it must consider the interaction between different indicators. For example, the coverage area is inversely proportional to the deployment cost. If the deployed solution with low cost is always chosen, the coverage area cannot be maximum. Therefore, the key point to achieve high deployment quality is to adjust the weight value between indicators as well as strike the balance. In other words, the evaluation mechanism can directly affect the accuracy. The process includes a series of processes that aim to achieve development goals. The design of the steps from the initial environment to the method will affect the effectiveness of the deployment strategy. Algorithms (methods)

usually use mathematical instructions or rules. A apposite algorithm can be adopted to avoid falling into local optima quickly; in other words, the algorithm is able to find the best deployment method for the global solution.

Key Applications

The deployment strategy of cellular network is widely applied. Most goals are based on minimum deployment cost or minimum energy consumption to achieve maximum coverage area, maximum number of service users, and maximum signal quality. In order to cover all the areas, Tseng et al. (2016) adopted macrocell, microcell, and femtocell to design deployment strategy. They proposed the top-down algorithm and bottom-up algorithm. The former aims to provide service to maximum coverage. The latter is to reduce the deployment cost. Furthermore, this study provides the best signal quality for each served user with low construction cost. Because the goal of this kind of problem is to provide Internet access anywhere, it will not take into account the number of users that can be serviced by RBS.

On the contrary, for the maximum number of service problems, the people within the coverage area of RBS cannot access the network likely because the service capacity of RBS is limited and such scenario usually happened in population centers. For maintaining the network performance, Fehske et al. (2009) introduced an evaluation mechanism for area power consumption and area throughput, through deploying a microcell to decrease the area power consumption in the network while still achieving certain area throughput targets. In other words, it can keep the number of service users and minimize energy consumption.

For the maximum signal quality strategy, it is necessary to consider the interference which is caused by obstacles. The obstacles include terrain, buildings, and so on. Klessig et al. (2011) introduced a traffic model that takes into consideration the co-channel interference and non-full load scenarios. Authors adjust the macro site

transmit power to fulfill the coverage condition of 95%, and the efficiency gain is also achieved by 20%.

The deployment strategy of cellular network will be more complex and difficult while network architecture changed. In order to extend the coverage, the concept of relay node has been proposed by Pabst (2004). It can reduce infrastructure deployment and enhance the capacity in cellular networks by using multichip communications via relay nodes.

To sum up, the application of deployment strategy of cellular network is wide and various. Many types of small cell and different network architectures lead to deployment problem and become a complex and important issue. Only providing a completed deployment strategy, the energy consumption and deployment cost can be decreased.

Cross-References

- [Base Station Cooperations Under Imperfect Conditions](#)
- [Energy Efficiency of Cellular Networks](#)
- [Heterogeneous Small Cell Networks](#)
- [Relaying in LTE-Advanced](#)

References

- Fehske AJ, Richter F, Fettweis GP (2009) Energy efficiency improvements through micro sites in cellular mobile radio networks. In: GLOBECOM workshops, pp 1–5
- Jump JR, Kirtane JS (1974) On the interconnection structure of cellular networks. *Inf Control* 24:74–91
- Klessig H, Fehske AJ, Fettweis GP (2011) Energy efficiency gains in interference-limited heterogeneous cellular mobile radio networks with random micro site deployment. In: Sarnoff symposium, pp 1–6
- Koutitas G, Karousos A, Tassiulas L (2012) Deployment strategies and energy efficiency of cellular networks. *IEEE Trans Wirel Commun* 11:2552–2563
- Pabst R et al (2004) Relay-based deployment concepts for wireless and mobile broadband radio. *IEEE Commun Mag* 42:80–89
- Tseng FH, Liang TT, Tassiulas L, Chou LD, Chao HC (2016) Network planning for heterogeneous cellular network in next generation mobile communications. *J Internet Technol (JIT)* 17:1269–1277

Device-to-Device (D2D) Communication

- [Resource Allocation in NOMA-Assisted D2D Networks](#)

Device-to-Device (D2D) Communications

- [Sidelink Transmissions and Proximity Services in LTE-A and Beyond](#)

Device-to-Device (D2D)-Assisted Cellular Systems

- [Resource Management in Macrocell–Small Cell Systems and D2D-Assisted Cellular Systems](#)

Device-to-Device Communication

- [Smart Device-to-Device Communication: Communication Establishment and Maintenance](#)

Device-to-Device Communications

- [Evolution of Vehicular Networks in the Transition from 3G to 4G Systems](#)

ϵ -Differential Indistinguishability

- [Differential Privacy](#)

ϵ -Differential Privacy

- [Differential Privacy](#)

Differential Confidentiality

- [Differential Privacy](#)

Differential Privacy

Meng Li
Beijing Institute of Technology, Beijing, China

Synonyms

Differential confidentiality; ϵ -Differential indistinguishability; ϵ -Differential privacy

Definition

Differential privacy is a mathematically formal definition of privacy which is used to quantitatively measure privacy loss.

Definition 1 (ϵ -differential privacy) A randomized function A satisfies ϵ -differential privacy (Dwork 2006; Dwork et al. 2006) if for all datasets D_1 and D_2 differing on at most one record, and all outputs $S \in \text{Range}(A)$,

$$Pr[A(D_1)S]e^\epsilon \times Pr[A(D_1) \in S], \quad (1)$$

where the possibility is taken over the randomness of A . The parameter $\epsilon > 0$ is a privacy budget. The smaller the ϵ is, the stronger the privacy protection is Dwork (2011) and Dwork and Roth (2014).

Historical Background

Differential privacy originates from database publication and privacy-preserving data analysis. People can initiate various queries on databases, such as health records, and cross-reference the results with query outputs from other databases to link or even identify a specific user. For example, a practical analysis (Narayanan and Shmatikov 2008) has been conducted on a Netflix dataset containing anonymized movie ratings of 500,000 Netflix subscribers, and it has shown that an adversary with a little background knowledge can easily identify a subscriber's record if the subscriber is present in the dataset. Therefore, a requirement that data records after a database publication cannot be traced back to their owners emerges. Differential privacy guarantees that it will not induce significant difference on the query result whether a data owner is in the database, so the adversary cannot obtain useful information about the data owner. Meanwhile, it gives a formal definition on privacy and how to quantitatively define the privacy level.

Foundations

A key notion of differential privacy is the sensitivity (Li et al. 2017) of a function f .

Definition 2 (Sensitivity) For any function $f: D \rightarrow R^d$, the sensitivity Δf of function f is:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \quad (2)$$

for all D_1 and D_2 differing on at most one record.

When $d = 1$, Δf is the maximum of the difference in the outcomes that f can query on two neighboring databases. For instance, if f is

a counting function, then adding or deleting a single database record can affect the outcome by at most 1 which means $\Delta f = 1$.

Two known approaches to achieve differential privacy are Laplace mechanism (Dwork et al. 2006) and exponential mechanism. Laplace mechanism is designed for functions whose outputs are real. It takes as inputs a database D , a function f , and a privacy budget ϵ . Specifically, a Laplace noise is sampled and added to the output. The Laplace noise is drawn from a Laplace distribution $Lap(b)$ with a probability density function $Pr[x|\mu, b] = \frac{1}{2b}e^{-(|x - \mu|)/b}$, where μ is the location parameter and b is the scale parameter determined by $b = \Delta f/\epsilon$. Exponential mechanism (McSherry and Talwar 2007) is designed for functions whose outputs are not real. It defines a utility function U which matches a utility score to each output and assigns higher probabilities to be selected exponentially to outputs with higher utility scores.

When taken into joint analysis, a sequence of functions should still provide privacy guarantee. There are two properties (McSherry 2009) of differential privacy regarding composition: sequential composition and parallel composition. First, there are several functions, and each function gives ϵ -differential privacy separately. A sequence of these functions gives $\sum_i \epsilon_i$ -differential privacy; second, if the sequence of functions is conducted on disjoint databases, then the final privacy budget is not an accumulated value but determined by the biggest privacy budget.

Key Applications

Differentially private distributed online learning for wireless big data. In a scenario where many sensor nodes are deployed in a rain forest and each node senses the environmental data, i.e., air temperature (Li et al. 2015) and precipitation. These data are locally processed to predict the rainfall because uploading this large scale of data will incur heavy communication overhead for typical wireless sensor networks. When the nodes are making predictions with online learning and

distributed convex optimization, they have to exchange weighted parameters with neighboring nodes at each round and update local parameters. However, since the nodes are subject to capture attack in a wild and unattended operation area, communications among nodes may leak the localization of nodes. A random Laplace noise can be added to the local parameter to guarantee differential privacy which is also output perturbation. To offset the prediction error caused by partial data and noises, a regret of the algorithm is defined to achieve a faster convergence rate.

Differentially private wireless data publication. Wireless local area networks have been used to evaluate network performance and analyze the user activity pattern since they trace association and disassociation records between users' wireless devices and the networks' access points (Yu et al. 2015). However, these records contain users' sensitive information, such as location and trajectory. Publishing these records without any protection will violate their privacy. Therefore, it is crucial to preprocess users' sensitive information before they are published. The other challenge is that the data utility for data analysis should be guaranteed without conflicting with differential privacy on multidimensional datasets. Suppose a trace database is a table T with w attributes, and it is associated with a joint frequency array A_T recording the statistical information of each attribute in table T . In a sanitization algorithm, a compact synopsis S_T of A_T is built by using multidimensional wavelet transforms. Then, a Laplace noise is added to each element in S_T to generate a noisy array $\hat{S}_T$. Lastly, the reconstruction information is recorded for each element in $\hat{S}_T$ to support query functions. Here, the supported query functions include Count and Sum.

Location privacy based on differential privacy in industrial wireless sensor networks. Industrial wireless sensor networks (IWSNs) are becoming an important component in Industry 4.0. A large number of moving nodes, such as workers and mobile products, are utilized to provide flexibility and mobility. These nodes upload sensing data to a data center to enable process monitoring and decision-making (Wang

et al. 2018). However, data mining technology will bring up the problem of privacy leakage because the data center is considered to be an honest-but-curious entity. For instance, a 3-month study (De Montjoye et al. 2015) on credit card records demonstrated that four spatiotemporal points were enough to recognize 90% of individuals. Meanwhile, how to determine an optimal partition granularity and improve the data utility of location data are other problems for partition schemes based on quadtree or kd-tree. Suppose a data domain is split into $l \times l$ cells $\{c_1, c_2, \dots, c_{l \times l}\}$ and each point number x_i in cell c_i is added by a random Laplace noise. Then, the sanitized database $\hat{\mathcal{D}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{l \times l}\}$ is released. Given a location database $\mathcal{D}$ and a privacy budget ϵ , how to improve the data utility comes down to choosing an optimal partition granularity g to achieve a minimal error $\min_g(e_n + e_u)$ of the query function where e_n is the noise error caused by noises and e_u is the nonuniformity error caused by the assumption of uniform distribution. Specifically, g is inferred through the linear-regression analysis, and a granularity partitioning model can be established. Second, a uniform cell release approach is constructed based on the granularity partitioning model. Then, similar cells will be merged into one cell if their hash value difference is smaller than a threshold and a random noise is added to each new cell. Lastly, the perturbed database $\mathcal{D}$ is released for query functions.

Cross-References

► [Differential Confidentiality](#)

References

- De Montjoye YA, Radaelli L, Singh VK, Pentland AS (2015) Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* 347(6221): 536–539
- Dwork C (2006) Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I (eds) *ICALP 2006*. LNCS, vol 4052. Springer, Heidelberg, pp 1–12
- Dwork C (2011) A firm foundation for private data analysis. *Commun ACM* 54(1):86–95

- Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. *Found Trends Theor Comput Sci* 9(3–40):211–407
- Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: *Proceedings of TCC*, pp 265–284
- Li CC, Zhou P, Jiang T (2015) Differential privacy and distributed online learning for wireless big data. In: *Proceedings of the WCSP*
- Li M, Zhu L, Zhang Z, Xu R (2017) Achieving differential privacy of trajectory data publishing in participatory sensing. *Info Sci* 400–401:1–13
- McSherry FD (2009) Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: *ACM SIGMOD*, pp 19–30
- McSherry F, Talwar K (2007) Mechanism design via differential privacy. In: *Proceedings of the FOCS*
- Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: *Proceedings of IEEE S&P 2008*, pp 111–125
- Wang J, Zhu R, Liu S, Cai Z (2018) Node location privacy protection based on differentially private grids in industrial wireless sensor networks. *Sensors* 18(2):1–15
- Yu J, Dong X, Luo Y, Li ML (2015) Differentially private wireless data publication in large-scale WLAN networks. In: *Proceedings of IEEE ICPADS*

Diffusion

- [Brownian Motion](#)

Digital Pseudonyms

- [Mixed Network](#)

Digital Right Management

- [Multimedia Security](#)

Direct Transmissions

- [Sidelink Transmissions and Proximity Services in LTE-A and Beyond](#)

Distance Distribution

- [Random Distance Distribution](#)

Distributed Anchoring

- [Distributed IP Mobility Management](#)

Distributed Beamforming in Wireless Sensor Networks

- [Collaborative Beamforming in Wireless Sensor Networks](#)

Distributed Clustering Algorithm

- [Distributed Community Detection in Opportunistic Networks](#)

Distributed Community Detection in Opportunistic Networks

Junling Shi and Xingwei Wang
College of Computer Science and Engineering,
Northeastern University, Shenyang, China

Synonyms

[Distributed clustering algorithm](#); [Distributed domain division algorithm](#); [Distributed grouping algorithm](#)

This work is supported by the National Natural Science Foundation of China under Grant No. 61572123; the Program for Liaoning Innovative Research Term in University under Grant No. LT2016007.

Definition

A community is a clustering of entities that are closely linked to each other, by either direct link-ages or easily accessible entities that act as intermediates. The distributed community detection algorithm can find such community structures in a fully distributed opportunistic network.

Historical Background

At present, people widely use mobile devices to generate and consume digital media contents, producing large volumes of traffic. By 2020, the number of mobile users is expected to reach 70% of the global population (CISCO 2016). When there are no infrastructures and content providers in the network, mobile devices (nodes) can only communicate with each other via a wireless connection which may break suddenly, which is the Delay Tolerant Network (DTN). In DTN, nodes are strongly interdependent but human-centric because they contact with each other following the communication pattern of human beings, such as Mobile Social Network (MSN) and Pocket Switch Network (PSN). This phenomenon motivates researchers to borrow the concept of Social Network Analysis (SNA) to design useful schemes. As one of the most leveraged SNA, community detection is widely used.

A community is a set of nodes with a high density of internal links, whereas links between communities have comparatively lower density (Katsaros et al. 2010). In the research community, it is widely believed that identifying community information about recipients can help select suitable forwarders and reduce the delivery cost compared to naive oblivious flooding. It is a reasonable intuition because people in the same community are likely to meet each other regularly and to be appropriate forwarders of messages destined for other members in the community (Hui et al. 2007).

Inspired by the above standpoint, when the community structure is introduced into DTN, the following benefits can be found. On one hand,

with the community structures, a message is forwarded to the specific detected community and then to the destination node. It enhances the DTN routing effectiveness because the forwarding destination is extended as a group of nodes rather than just one destination node, which makes packet delivery be achieved easier and faster. On the other hand, social ties among community members are more entrenched than the social relationship between two nodes; therefore, the community-aware routing mechanism is relatively stable compared with routing without community structures.

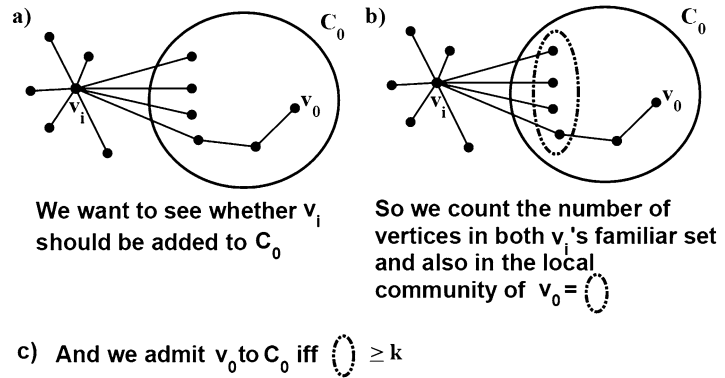
In sociology, most community detection approaches are centralized, that is, the detection results are obtained based on the overall view of the network. However, in DTN, obtaining such an overall view of the network is unrealistic since nodes are fully distributed. Therefore, distributed community detection approaches should be designed. One of the famous community detections is the distributed k -clique community detection algorithm (Hui et al. 2007). A node records the encounter node as its community member if the latter has long encounter duration or many common friends with the former. The brief detection process is shown in Fig. 1 (Hui et al. 2007).

Firstly, each node (assume as v_0) needs to maintain the following information: a list of encountered nodes and the contact durations with each of them, the familiar set F_0 , the local community C_0 , and local approximation of the familiar sets of all nodes in its local community C_0 , which is denoted as $FSoLC_0$, that is

$$FSoLC_0 = \{\tilde{F}_j | v_j \in C_0, j \neq 0\} \quad (1)$$

Next, when v_0 encounters another v_i ($i \neq 0$), if v_i is not in F_0 and their total contact duration count exceeds a certain threshold, v_0 will insert v_i into F_0 and C_0 . If v_i is not in C_0 and F_i contains at least $k - 1$ members of C_0 , then v_0 adds v_i into C_0 . Besides, if $\tilde{F}_j$ contains at least $k - 1$ members of C_0 , v_j is added to the local community C_0 . Based on the simulation, it is observed that with a suitable threshold k , the k -clique can reach

Distributed Community Detection in Opportunistic Networks,
Fig. 1 The main principle of k -clique community detection (Hui et al. 2007)



around 80% of the performance of centralized algorithm.

Key Applications

MSN

MSN is a mobile communication system which involves social relationships of users. In such network, mobile users can access, share, and distribute data (e.g., video) in a mobile environment by exploiting their social relationships (Vastardis and Yang 2013). MSN has three architectures, i.e., the centralized one, the distributed one, and the hybrid one (Kayastha et al. 2011). In a distributed architecture, nodes connect with each other via the opportunistic encounters, which makes the communication mode become a DTN. Benefit from the detected community structures, MSN nodes are able to share their resources with the community members and route messages among each other more efficiently.

VANET

As a new technology that integrates the potentials of new-generation wireless networks into vehicles to enable communication among vehicles via Mobile Ad hoc Network (MANET) (Bonola et al. 2016), VANET (Vehicular Ad-hoc Network) is attractive to be integrated into smart cities. VANET is distinguished from other kinds of ad hoc networks due to its rapidly changing topology, large scale, and variable node

density. Generally, VANET can be classified into Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V); the former is the communication mode via infrastructures and the latter is that via vehicles only (Al-Sultan et al. 2014). VANET is a kind of DTN when it is established as the latter mode. Since VANET is formed by human-centric nodes (vehicles), community structures can also be employed to improve the VANET routing efficiency.

In-Network Caching

In-network caching allows nodes to cache the passed content locally. Thanks to the cached contents, forthcoming content requests can be responded in time. By detecting the community structures, user clusters are taken into account, which makes the network turn into some domains. It can improve the network scalability since nodes are autonomous, and a cluster head can be selected as the administrator to manage the community members, e.g., cache the popular data for the whole community. Therefore, the caching content of nodes can be rationally distributed.

Cross-References

- [Applications and Architectures of Social Internet of Things](#)
- [Cache Placement and Content Delivery Design in Wireless Caching Networks](#)
- [Capacity of Wireless Ad Hoc Networks](#)

References

- Al-Sultan S, Al-Doori MM, Al-Bayatti AH, Zedan H (2014) A comprehensive survey on vehicular ad hoc network. *J Netw Comput Appl* 37(1):380–392
- Bonola M, Bracciale L, Loretì P, Amici R (2016) Opportunistic communication in smart city: experimental insight with small-scale taxi fleets as data carriers. *Ad Hoc Netw* 43:43–55
- CISCO (2016) Cisco visual networking index: global mobile data traffic forecast update 2015–2020, San Jose, White Paper
- Hui P, Yoneki E, Chan SY, Crowcroft J (2007) Distributed community detection in delay tolerant networks. In: *Proceedings of 2nd ACM/IEEE international workshop on mobility in the evolving internet architecture*, pp 29–38
- Katsaros D, Dimokas N, Tassiulas L (2010) Social network analysis concepts in the design of wireless ad hoc network protocols. *IEEE Netw* 24(6):23–29
- Kayastha N, Niyato D, Wang P, Hossain E (2011) Applications, architectures, and protocol design issues for mobile social networks: a survey. *Proc IEEE* 99(12):2125–2129
- Vastardis N, Yang K (2013) Mobile social networks: architectures, social properties, and key research challenges. *IEEE Commun Surv Tutor* 15(3):1355–1371

Distributed Computing

- [Data-Driven Resource Allocation](#)

Distributed Data Security

- [Distributed Storage Security](#)

Distributed Domain Division Algorithm

- [Distributed Community Detection in Opportunistic Networks](#)

Distributed Grouping Algorithm

- [Distributed Community Detection in Opportunistic Networks](#)

Distributed Inference with Wireless Byzantine Data

- [Byzantine Attack and Defense in Wireless Data Fusion](#)

Distributed IP Mobility Management

Carlos J. Bernardos

Universidad Carlos III de Madrid, Madrid, Spain

Synonyms

[Distributed anchoring](#); [DMM](#)

Definitions

Distributed IP Mobility Management (DMM) refers to the concept of placing mobility anchors closer to the user, distributing the control and data infrastructures among the entities located at the edge of the access network. This allows the mobile node's traffic to take an optimal route while enabling the mobile node move between mobility anchors. DMM appears as an alternative to classical centralized IP mobility management approaches, complementing them for better performance and scalability.

Motivation

Classical IP mobility management solutions, such as Mobile IPv6 (Perkins et al. 2011) or Proxy Mobile IPv6 (Gundavelli et al. 2008), require the presence of a central entity (home agent or local mobility anchor) that anchors the IP address used by the mobile node (MN). This centralized function is in charge of coordinating the mobility management. As a result, these centralized mobility anchors are burdened with data forwarding and control mechanisms for a huge number of users. While this centralized way of addressing mobility management has been fully developed by the Mobile IP protocol family and its many extensions, it brings several limitations that have been identified in Liu et al. (2015):

- Suboptimal routing: Since the (home) address used by an MN is anchored at the home link, traffic always traverses the central anchor, leading to paths that are, in general, longer than the direct one between the MN and its communication peer. This is exacerbated with the current trend in which content providers push their data to the edge of the network, as close as possible to the users, as, for example, deploying content delivery networks (CDNs). With centralized mobility management approaches, user traffic will always need to go first to the home network and then to the actual content source, sometimes adding unnecessary delay and wasting operator resources.
- Scalability problems: Existing mobile networks have to be dimensioned to support all the traffic traversing the central anchors. This poses several scalability and network design problems, as central mobility anchors need to have enough processing and routing capabilities to be able to deal with all the users' traffic simultaneously. Additionally, the entire operator's network needs to be dimensioned to be able to cope with all the users' traffic.
- Reliability: Centralized solutions share the problem of being more prone to reliability

problems, as the central entity is potentially a single point of failure.

In order to address these issues, distributed IP mobility management approaches are being explored and standardized by the DMM working group at the IETF.

DMM Solutions

There exist several approaches that can be followed to design a DMM solution. While no solution has been standardized so far (the IETF DMM WG is working on it), proposals span from extensions of current standardized protocols to clean-slate solutions. Next, the operation of two main families of DMM approaches is summarized.

Mobile IPv6-Based DMM Solution

Mobile IPv6 solutions, like Giust et al. (2015), are based on deploying multiple home agents at the edge of the network. Multiple of these anchors can be simultaneously used by a single MN. When a mobile node attaches to a distributed anchor router, it gets an IPv6 address which is topologically anchored at that router. While attached to this router, the mobile can send and receive traffic without traversing any tunneling nor special packet handling. If the MN moves and attaches to a different anchor router, it gets a new IPv6 address from that router. In case the MN wants to keep the reachability of the IPv6 address(es) it obtained from previous anchor (e.g., note that this decision is dynamic may be done on an application basis), the host has to involve its Mobile IPv6 stack, by sending a Binding Update to the router where the IPv6 address is anchored, using the address obtained from the current anchor as care-of address.

In this way, the IPv6 address that the node wants to maintain in use plays the role of home address, and the anchor from where that address was configured plays the role of home agent (for that particular address). In this scenario, old flows are anchored to the previously visited router, which is in charge to encapsulate the packets

and deliver them to the MN's care-of address (CoA). The IP tunnel is bi-directional, so the MN does the same when sending packets with the old address. Conversely, new IP flows are started using the address configured at the new anchor. These flows are handled by the new distributed anchor as a plain IPv6 router (see Fig. 1).

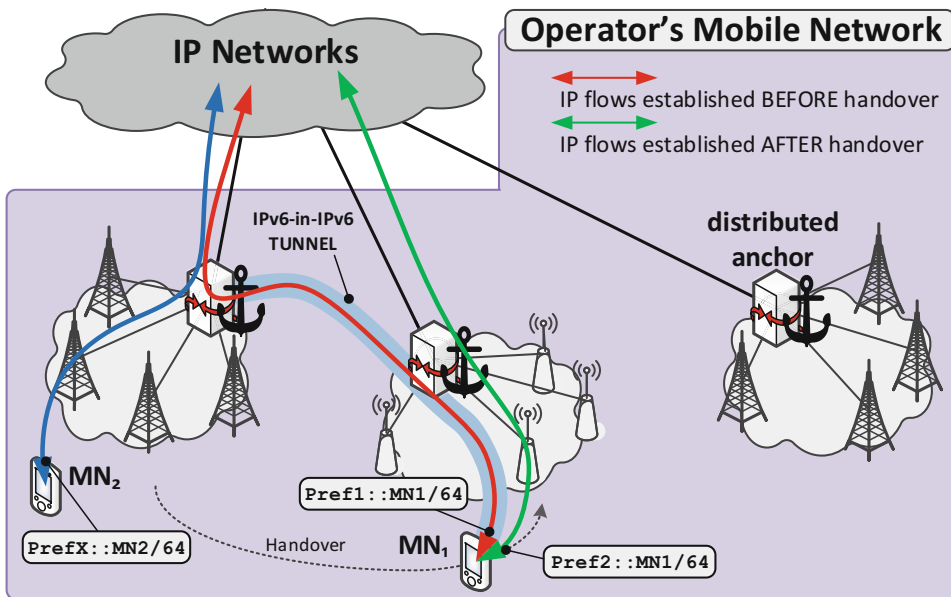
Proxy Mobile IPv6-Based DMM Solution

Proxy Mobile IPv6 solutions, like Giust et al. (2014), are based on replacing the role of the Mobile Access Gateway (MAG) by a new distributed anchor. This distributed anchor is provided with links to the Internet that do not imply paths traversing a Local Mobility Anchor (LMA). Hence, the distributed anchor acts as a plain access router (i.e., no tunneling) to forward packets to and from the Internet. Also, the distributed anchor features mobility anchoring functions, being able to forward without disruption the IP flows that an MN started while attached to it before moving to a new anchor afterward. The control plane functions of the LMA can be assumed by a control-only entity which stores, for every MN, all the

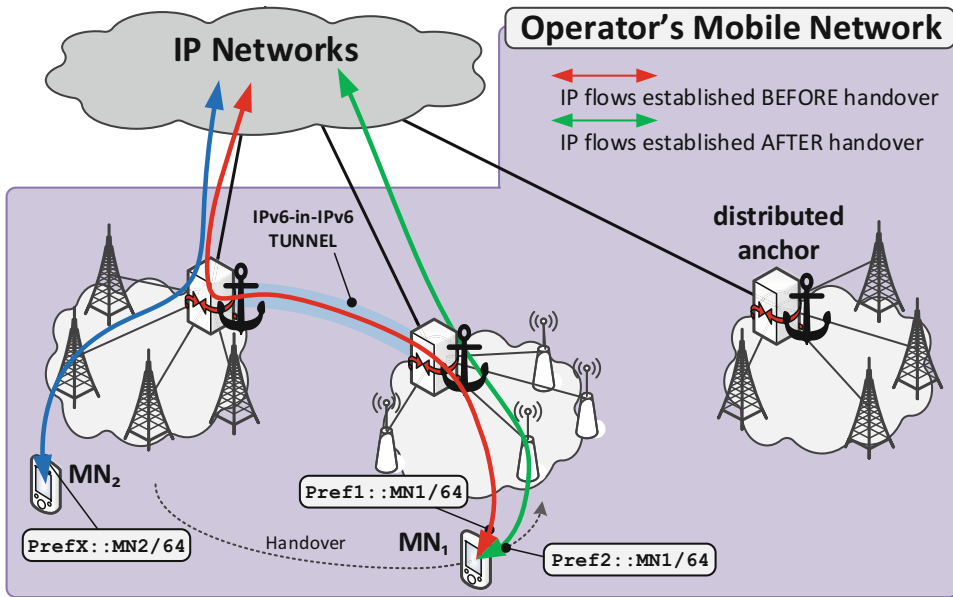
prefixes advertised to the MN, which anchor the MN is currently connected. This control plane entity is responsible of receiving/sending all the required mobility signaling (i.e., proxy binding updates and proxy binding acknowledgments) from/to the distributed anchors, so they can set up the required bi-directional tunnels to guarantee reachability of the addresses used by the MN, regardless of where these addresses are anchored (see Fig. 2).

Future Directions

DMM solutions are currently being standardized at the IETF. The 3GPP has also adopted distributed IP mobility management concepts in the 5G specifications of the cellular architecture (i.e., Release 15 and onward). In addition to the approaches described in this article, other solutions are also being explored, as the use of Segment Routing (Filsfilis and Previdi 2017), which might open the door for new ways of enabling network programmability.



Distributed IP Mobility Management, Fig. 1 MIPv6-based DMM solution



Distributed IP Mobility Management, Fig. 2 PMIPv6-based DMM solution

Cross-References

- [Mobile IP](#)
- [Network-Layer Mobility Management](#)
- [Proxy Mobile IPv6](#)

References

- Filsfils C, Previdi S (2017) Segment routing architecture. draft-ietf-spring-segment-routing-14
- Giust F, Bernardos CJ, de la Oliva A (2014) Analytic evaluation and experimental validation of a network-based ipv6 distributed mobility management solution. IEEE Trans Mobile Comput 13(11):2484–2497. <https://doi.org/10.1109/TMC.2014.2307304>
- Giust F, Bernardos CJ, De la Oliva A (2015) Hdmm: deploying client and network-based distributed mobility management. Telecommun Syst 59(2):247–270
- Gundavelli S, Leung K, Devarapalli V, Chowdhury K, Patil B (2008) Proxy mobile IPv6. RFC 5213
- Liu D, Zuniga J, Seite P, Chan H, Bernardos C (2015) Distributed mobility management: current practices and gap analysis. RFC 7429
- Perkins C, Johnson D, Arkko J (2011) Mobility support in IPv6. RFC 6275

Distributed Storage Security

Dengzhi Liu

Nanjing University of Information Science and Technology, Nanjing, China

Synonyms

[Cloud storage security](#); [Distributed data security](#); [Wireless data security](#)

Definitions

A distributed system divides end users' data into multiple segments. Each of partial nodes in the distributed system stores one or multiple data segments. However, the data segments stored in the distributed system may be tampered with by adversaries. Moreover, it is possible that the distributed system may discard some rarely accessed data segments to save the storage resource or alle-

viate some resource-constrained nodes' storage burden. Hence, the data storage security in the distributed system needs to be ensured.

Historical Background

The original intention of developing distributed computing is to use the idle computers to deal with other incidental computing tasks. Different from centralized computing, distributed computing can be used to process and store massive data with high efficiency and local resource saving. The first prototype of distributed computing was put forward by the Intel Corp at the end of the 1980s. With the explosive growth of Internet access terminals in the late 1990s and early twentieth century, the development of distributed computing technology has reached a climax. Currently, there are many information technologies developed by distributed computing, such as Mobile Agent technology, P2P technology, and Grid technology.

Cloud computing is a representative technology that is developed from distributed computing. Many distributed servers are connected together by the Internet in cloud computing. The cloud is managed by a cloud service provider (CSP). Users can enjoy the distributed storage services via the Internet. However, the data stored in the distributed servers may be tampered with and discarded by the CSP for financial reasons. According to a recent report by Twinstrata, only 50% of users choose the cloud to back up their data, and only 20% of them store their sensitive data in the cloud. Hence, how to protect the data in the distributed storage system is a serious issue that should be solved well. To avoid the data leakage, a cryptography method is used to encrypt the data before outsourcing it. A new emerging problem is how to verify the storage of the encrypted data in the distributed system.

Foundations

To ensure the distributed storage security, the system should allow users to check the integrity of

the stored data. Because users lose local control of their data after the data are outsourced to the remote distributed servers, they hope to check their stored data in the cloud at any time. In existing studies, there are three kinds of remote storage checking, which are described in the following.

- **Provable Data Possession (PDP)** PDP allows users to check whether their data stores in the remote server without retrieving all the data to the local side. The related studies can be found in Ateniese et al. (2007, 2008, 2011). Note that PDP is a probabilistic storage proof protocol. In other words, the user side can sample a random set of data segments and request the remote server to prove whether these data segments are stored in the corresponding server.
- **Proof of Retrievability (POR)** Similar to PDP, POR can also let users check the data possession in the remote server. Specially, POR protocols can allow users to check whether the data stored in the remote server can be retrieved to the local side (Juels and Kaliski 2007; Bowers et al. 2009; Dodis et al. 2009). In POR, a common approach is that the user generates authenticators for every data segments before data outsourcing. The authenticators can be used for the further verification of the data possession and retrievability.
- **Storage Auditing** In general, the data storage verification result may determine the reputation level of the remote storage server. If the data stored in one server is often destroyed, the number of users who choose to use the corresponding server to store their data will decrease. Moreover, the computational cost of data storage checking is also great. To alleviate the burden of the local user side and improve the fairness of the data storage checking, some researchers resorted to a TPA (third-party auditor) to design the data storage auditing protocols based on the homomorphic authenticator (Wang et al. 2010, 2013, 2011). In the storage auditing protocols, the user computes data tags and signatures for data segments and then outsources them with data segments to the server. Note that the TPA can

audit the data storage on behalf of the users without knowing the details of the stored data.

A comparison of the properties that the three protocols can provide is shown in Table 1.

Applications in Wireless Networks

Wireless terminals can communicate with each other via Wi-Fi or cellular network in wireless networks. Similar to the construction of the distributed system, end users can resort to neighboring wireless terminals to compute and store their data. Compared to nodes in the traditional distributed system, the wireless terminals have lower resource of the computing and the storage. Moreover, the information transmitted over wireless channels is more easily captured by attackers. Hence, the communication protocols in the wireless networks should be more secure, and more terminals need to be aggregated in the

wireless distributed system. A typical wireless distributed system is shown in Fig. 1. Note that there are many lightweight wireless devices in the wireless distributed system. These devices can communicate each other via wireless networks, and the aggregated data are forwarded to end users by remote wireless transmission infrastructures. Some applications of distributed data storage in wireless networks are listed in the following.

- *Secure distributed data storage for wireless networks* Wireless end users can eliminate the disadvantage of limited storage space by storing data segments in surrounding wireless terminals over wireless networks.
- *Secure distributed data computing for wireless networks* Wireless end users can delegate complex computing tasks to nearby wireless terminals whose computing capability is powerful than their own.
- *Secure distributed data sharing for wireless networks* Wireless terminals can transmit their own information to end users through wireless networks, and end users can also transmit their own data to any wireless terminal in the system through wireless networks.

Distributed Storage Security, Table 1 Properties of various data storage checking protocols

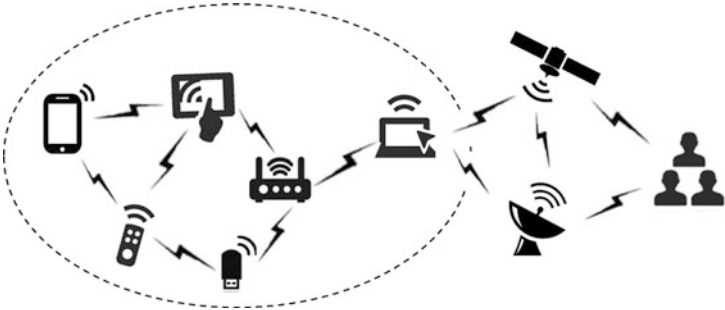
Properties	PDP	POR	Auditing
Publicly checking	N	N	Y
Local side checking	Y	Y	N
Data dynamics	Y/N	Y	Y
Batch verification	N	N	Y

Y: The protocol can provide the corresponding property
N: The protocol cannot provide the corresponding property
Y/N: The protocol cannot fully provide the corresponding property

Distributed Storage Security, Fig. 1 A typical distributed system for wireless networks

Future Directions

In the research of the distributed storage security, some future directions are summarized as follows.



- *More Efficient Data Dynamics* More efficient data dynamics implies that the distributed storage assurance protocols should support efficient data dynamics with low computational cost and high accuracy.
- *Data Deduplication* In a distributed storage system, it is possible that the same data block may be stored in two or more nodes, which wastes storage resource and may affect the verification of the data storage. Hence, it is necessary to propose some secure and effective data deduplication protocols for the distributed system.
- *Corrupted Data Location* In the batch verification of the data storage, the fail of the verification may be caused by one data segment or part of the data segments. Hence, a protocol that can locate the corrupted data segments in the batch storage verification needs to be proposed.

Cross-References

- [Data-Driven Security](#)
- [Verifiable Cloud Computing](#)
- [Wireless Big Data](#)

References

- Ateniese G, Burns R, Curtmola R, Herring J, Kissner L, Peterson Z, Song D (2007) Provable data possession at untrusted stores. In: ACM conference on computer and communications security, pp 598–609
- Ateniese G, Pietro RD, Mancini LV, Tsudik G (2008) Scalable and efficient provable data possession. In: Proceedings of the international conference on security and privacy in communication networks, pp 1–10
- Ateniese G, Burns R, Curtmola R, Herring J, Khan O, Kissner L, Peterson Z, Song D (2011) Remote data checking using provable data possession. ACM Trans Inf Syst Secur 14(1):1–34
- Bowers KD, Juels A, Oprea A (2009) Proofs of retrievability: theory and implementation. In: ACM cloud computing security workshop, pp 43–54
- Dodis Y, Vadhan S, Wichs D (2009) Proofs of retrievability via hardness amplification. In: Theory of cryptography, Springer, Berlin, Heidelberg, San Francisco, CA, pp 109–127
- Juels A, Kaliski BS (2007) Pors: proofs of retrievability for large files. In: ACM conference on computer and communications security, pp 584–597
- Wang C, Wang Q, Ren K, Lou W (2010) Privacy-preserving public auditing for storage security in cloud computing. In: Proceedings of international conference on computer communications, pp 1–9
- Wang Q, Wang C, Ren K, Lou W, Li J (2011) Enabling public auditability and data dynamics for storage security in cloud computing. IEEE Trans Parall Distrib Syst 22(5):847–859
- Wang C, Chow SSM, Wang Q, Ren K, Lou W (2013) Privacy-preserving public auditing for secure cloud storage. IEEE Trans Comput 62(2):362–375

Distributed Transmission Coordination

- [Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs](#)

Diversity Forwarding

- [Opportunistic Routing \(OR\)](#)

Division Multiplexing

- [Multiple Access Technique for Cellular Wireless Networks](#)

DMM

- [Distributed IP Mobility Management](#)

Drug Delivery via Nanomachines

Yifan Chen¹, Yu Zhou^{2,3}, Neda Sharifi¹,
Ross Murch³, and Geoffrey Holmes¹

¹The University of Waikato, Hamilton, New Zealand

²Southern University of Science and Technology, Shenzhen, China

³The Hong Kong University of Science and Technology, Hong Kong, China

Synonyms

Targeted drug delivery

Definition

Drug delivery via nanomachines is one type of targeted drug delivery (TDD) technologies that delivers medicament to the designated area in vivo by using devices ranging in size from 0.1 to 10 μm and constructed of nanoscale components. Various terms such as nanomachine, nanorobot, nanobot, nanoid, nanite, and nanomite have been used to describe such devices. These nanomachines are loaded with drugs and targeted solely to diseased tissue under the manipulation and guidance of an external monitoring system. The purpose of this technology is to remarkably enhance locoregional therapies for disease treatment while reducing the side-effects of the drug on healthy cells.

Historical Background

In traditional drug delivery systems such as oral ingestion or intravascular injection, the medication is distributed throughout the body via the systemic blood circulation. For most therapeutic agents, only a small portion of the medication reaches the organ to be treated. Conventional chemotherapy shows that only a small fraction of drugs (0.1%) are taken up at the tumor cells,

whereas the remaining part (99.9%) are absorbed by healthy tissue (Trafton 2009).

On the other hand, TDD seeks to concentrate the medication in the site of interest while reducing the relative concentration of the medication in the remaining regions. For example, by avoiding the host's defense mechanisms and inhibiting nonspecific distribution in the liver and spleen (Bertrand and Leroux 2011), a system can reach the intended site of action in higher concentrations. This ability for drug molecules to concentrate in areas of diseased tissue is accomplished through three targeting mechanisms: passive, active, and physical. Passive targeting refers to producing a natural distribution difference of the drug in vivo by utilizing the specific physiological and structural characteristics of tissues and organs, thereby achieving a targeting effect. One example is the preferential accumulation of chemotherapeutic agents in solid tumors as a result of the difference in vascularization of tumor tissue compared with healthy tissue (Sagnella and Drummond 2012). Active targeting confers the drug an ability to actively bind to the target by attaching the probe molecules with chemical or physical methods, thereby enhancing the specificity of nanomachines to the targeted site. One example is to utilize cell-specific ligands that will allow for nanoparticles to bind specifically to cells that have the complementary receptor. Physical targeting uses physical signals such as light, heat, ultrasound, electric field, and magnetic field to artificially regulate the distribution and releasing of drugs in the body and achieve targeted delivery.

TDD via nanomachines is the technology that utilizes nanomachines to further improve its targeting efficiency. Similar to other research topics in nanotechnology, the origin of this technique comes from the Nobel Prize-winning theoretical physicist Richard Feynman, who first presented the idea of nanomachines in 1959. In his seminal speech "There's Plenty of Room at the Bottom" (Feynman 1992), Feynman envisioned that in the future humans might build a molecule-sized miniature machine to "arrange the atoms the way we want" and to perform chemical synthesis by mechanical manipulation. This led to the idea of

using nanorobots to treat diseases, which in Feynman's words is called "swallowing the doctor."

TDD via nanomachines relies heavily on the development of nanorobotic technologies. After decades of exploration, the research of manufacturing suitable nanorobots is mainly concentrated in two directions: (i) engineering the existing bacteria in nature with specific functions to create biological robots for drug delivery (Martel et al. 2009) and (ii) fabricating an artificial carrier with synthetic material that is harmless to the human body for drug delivery (Bae and Park 2011).

During the TDD process, there are several environmental uncertainties such as circulation and diffusion of nanomachines in the blood stream, which have a significant impact on the drug accumulation in the targeted site (Bae and Park 2011). In order to overcome these limitations and improve the transport efficiency, control strategies have been introduced to guide the movement of nanomachines (Liu et al. 2017; Guo et al. 2015). For instance, an optical-control mechanism proposed in Liu et al. (2017) generated dynamic tweezers to spin the nanomachines by a modified optical filter, which achieved bright and dark soliton conversion behaviors. Furthermore, a magnetic-control mechanism conveyed the device from one location to another along a predetermined path by utilizing the magnetic gradient propulsion. Another challenge during TDD is the biodegradability and biocompatibility of nanomachines. Injecting nanomachines into the human body for medical applications requires these devices to be biologically harmless, where the drug concentration should not reach toxic levels (Bae and Park 2011).

Communication models and theories have been applied to optimize TDD by modelling the procedure as an information sending and receiving process, where drug nanoparticles are information molecules through which messages are encoded, and injection and delivery of drugs correspond to the transmitting and receiving processes (Chahibi et al. 2015a; Chen et al. 2017a). This approach transforms TDD by applying performance measures traditionally only known to the communi-

cations community to rigorously define the efficiency of TDD. Furthermore, it allows for the utilization of classical communications techniques and protocols to design optimally targeted therapies. While most of the works do not consider utilizing externally controllable nanocarriers for TDD, Chen et al. (2017a) focus on the scenario of nanorobots-assisted TDD. Communications-inspired TDD via nanomachines has been incorporated in the first IEEE Standard on Nanoscale and Molecular Communications 1906.1 as a use case (IEEE NanoCom Standards Development Working Group 2017).

Foundations

Engineered bacterial nanomachines: The most widely considered bacterial nanomachines are the magnetotactic bacteria (MTB) with flagellated nanomotors combined with the nanometer-sized magnetosomes as a means of propulsion or bio-actuation (Martel et al. 2009). MTB synthesize intracellular Fe_3O_4 nanoparticles, called magnetosomes, assembled in a chain acting as a compass to guide their movement following the applied magnetic field. Engineered bacterial nanomachines have a size varying in the range of 0.5–3 μm (Martel et al. 2009). The advantage of using bacterial nanomachines is to avoid the difficulty of developing an autonomous (i.e., without the need for an external power source for propulsion) engine capable of sufficient thrust for practical applications. A swarm of MTB can be used to push, transport, and manipulate a microobject with precise control over their displacement (Martel et al. 2009). In addition to MTB, there are other bacterial materials that can be used to manufacture bio-nanomachines, such as cellulosome which are a complex of multicellulolytic enzymes that dismantle plant polysaccharides (Artzi et al. 2017).

Artificial nanomachines: The most commonly considered artificial nanomachines are built upon chemically synthesized simple geometric structures such as helical, tubular, and capsule shapes (Chen et al. 2017b). Due to the limitation

in fabrication, at present it is difficult to embed on-board power sources within these nanomachines; therefore, the magnetic field is also widely used for their actuation and control. However, unlike MTB which utilize their flagellated nanomotors to move forward, artificial nanomachines obtain their forward momentum under the influence of a magnetic field via spin. For instance, once a nanomachine is placed inside a uniform magnetic field generated by an approximate Helmholtz electromagnetic coil system, the magnetic field creates a torque on the nanomachine, causing it to rotate. Subsequently, the nanomachine maintains its movement as it heads towards the coil with its original orientation (Kim et al. 2016). Most nanomachines under development have characteristic lengths in the order of $1\text{ }\mu\text{m}$ to $100\text{ }\mu\text{m}$ and are fabricated using a wide variety of techniques including direct laser writing, templated directed electrodeposition, self-scrolling, shadow-growth, microfabrication, underpotential deposition, to name a few (Kim et al. 2016).

Drug loading and releasing: Before being injected into the human body, drug molecules are attached to the surface of a nanomachine by chemical bonding (Hu et al. 2014) or are encapsulated inside a nanomachine (Chen et al. 2017b). Drug releasing in vivo is a very challenging task. A common strategy is to use external stimuli to make the outer package degrade and cause the release of the encapsulated drug. Such stimuli include pH, temperature, light, ultrasound, magnetic field, and enzymes. For instance, when a magnetic liposome is applied to nanomachine manufacturing, it will incorporate degradable polymers, which can degenerate under an external stimulus and release the drug at a specified time (Chude-Okonkwo et al. 2016). Another example is to utilize oxidoreductases as a gate in the responsive controlled release system. Glucose oxidase is extensively used as a diagnostic tool in the detection of glucose values. Upon addition of the hyperglycemic state, glucose oxidase encapsulated in the nanomachine begin converting glucose into gluconic acid, resulting in dissociation of the package and subsequent release of insulin (Hu et al. 2014).

Controlling mechanism: Major advances have been achieved to control the motion of nanomachines to steer them towards desired sites by inducing external gradients. These stimuli are categorized as chemical, optical, ultrasonic, electric, and magnetic.

Chemical stimuli can trigger swarming behavior of different types of catalytic nanomachines ranging from simple to more complex units such as Janus spheres and Au nanowires. For instance, chemically induced aggregate behavior has been observed in self-propelled Au nanoparticles. The process contains chemical reaction between hydrazine and H_2O_2 catalyzed by the gold surface that generates chemical gradients on the nearby Au nanoparticles, which pull the particles towards the center of the swarms where the largest electrolyte gradient existed (Liu et al. 2017). However, for chemically triggered swarming behavior, the swarming time is highly dependent on the reaction speed, which limits its applications.

Optical stimuli have also been an efficient method for inducing aggregation of nanomachines. This mechanism is defined as the use of intensely focused light beams to accurately trap and manipulate nanomachines. Intensity gradients generated by a converging beam are able to polarize nanomachines to move towards the highest gradient region of the electric field (Liu et al. 2017).

Another effective way to regulate the collective behavior of nanomachines is ultrasound, which is an on-demand motion control, long lifetime, and noninvasive aggregation method. Interactions between ultrasound and nanomachines lead to swarming phenomena based on pressure nodes or antinodes, where the nanomachines migrate from peaks and accumulate in the troughs of the associated standing waves (Guo et al. 2015).

Electric stimuli induce dielectrophoresis, which is a phenomenon exerted on dielectric nanomachines suspended in water solution to initiate the gathering of nanomachines (Liu et al. 2017). The application of this method to TDD has been limited because it requires complex configuration to generate appropriate

field and high voltages that may damage tissue.

Magnetic stimuli are the most common way to remotely control the motion of nanomachines. Magnetic fields can be used to manipulate the motion of nanomachines as well as track their movement. This stimulus also has the capability of superfast operation as its response time is just a few seconds. Injected nanomachines loaded with drug molecules can be aggregated and guided by a local magnetic field, which can be applied in nearly all human tissues without any toxicity (Liu et al. 2017). Table 1 gives a comparison between several different types of nanomachines and their control mechanisms.

Communication-inspired system design: The molecular communication (MC) models can be applied to describe the TDD process (Chahibi et al. 2015a; Chen et al. 2017a). In MC, an engineered miniature transmitter emits molecules, which propagate and are eventually received by a miniature receiver. The information may be encoded in either the timing or the concentration of molecules. The reception process often involves a chemical reaction between molecules (ligand) and compliant receptors present on the receiver surface.

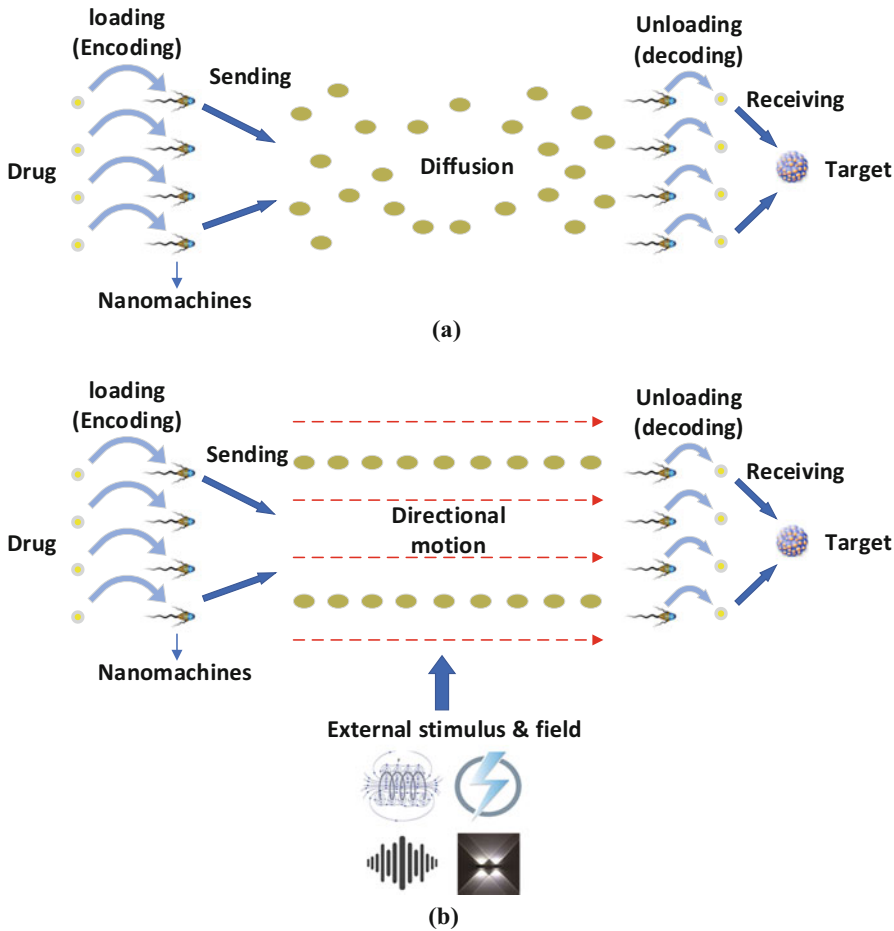
Under this framework, the drug nanoparticles are the information molecules through which messages are encoded, the injection and delivery of the therapeutic agent correspond to the transmission and reception of signals, and the blood vessels responsible for the transport of the drug are the communication channel. The signal is the concentration-time profile of the therapeutic agent, which should be successfully received by the tumor to decrease its size. Figure 1 provides general molecular communication system models with and without external controlling factor.

This analogy enables the analysis and design of TDD by utilizing performance measures traditionally only used in the communication field. By applying the queuing theory to model receptor saturation, thus treating it as a congestion problem in communication networks, the optimal drug release rate was estimated in Femminella et al. (2015) to guarantee that the desired fraction of receptors bound to drug molecules without drug over loading. By modeling the pharmacokinetics of TDD systems as the channel impulse response in communication systems, the optimal drug injection rate was obtained in (Chahibi et al. 2015a). By applying the antibody-mediated TDD system transport and antigen-binding kinetics and taking into account the geometry of the antibody molecule, the electrochemical structure of the antibody-antigen complex, and the physiology of the patient, optimal shape of the antibody molecular structure was determined in (Chahibi et al. 2015b). By using contrast-enhanced microwave imaging as a pretherapeutic evaluation technique to determine the pharmacokinetic model of nanorobots-assisted TDD, strategies for optimal targeted therapies were presented from the communication waveform design perspective (Chen et al. 2017a).

Vascular network channel model: Due to the limitations of modern medical imaging technologies, it is difficult to obtain the exact values of all important vascular parameters such as topology, size, orientation, speed of blood flow. Thus, in order to analyze the pharmacokinetics of TDD systems through the abstraction of MC, it is necessary to utilize statistical approaches to simulate the vascular morphology. The fractal model is often chosen as the first approximation because the distributive function of vasculature is to bring

Drug Delivery via Nanomachines, Table 1 Several types of nanomachines and corresponding control mechanisms

Type	Controlling mechanism	Propulsion
Magnetotactic bacteria	Magnetic field	Flagella
Magnetic nanoparticles	Magnetic field	Spin
Plasmonic motor	Light	Light-induced rotation
Chemotaxis nanomotor	Chemical gradient	Chemical reaction
Electric tweezers/nanowires	Electric field	Electrophoretic and dielectrophoretic forces



Drug Delivery via Nanomachines, Fig. 1 Communications-inspired drug delivery system model: (a) without an external controlling unit and (b) with a controlling unit

a stream of blood to every living cell within the body or organ and is achieved by successive dichotomous division, or bifurcation. The self-similarity of bifurcation at each node gives the vascular connection a fractal character. In Chen et al. (2017a), a fractal-based model is proposed to consider the random realization of a single propagation path on the fractal tree from the transmitter (injection site) to the receiver (cancer cells). Another similar model in Chahibi et al. (2015a) introduces an interesting method that maps the cardiovascular network into an equivalent circuit. Meanwhile, a more detailed model based on Chen et al. (2017a) is presented in Zhou et al. (2018) to explore the properties of the grid-like region near the tumor site.

Key Applications

Drug delivery via nanomachines is the most promising solution for TDD. Drug-loaded nanomachines are notably utilized in different environments and operations relevant to the treatment and prevention of diseases such as metabolic disorders, inflammatory diseases, autoimmune diseases, neurodegenerative diseases, and cancer. Among them, the application of cancer treatment has drawn the most attention. The pain caused by conventional cancer treatments such as chemotherapy that are often accompanied by huge side effects is not even better than the pain caused by cancer itself. Most of the drugs applied in the treatment of cancer

have a certain toxicity (Bae and Park 2011). This toxicity not only kills healthy cells but also affects the size of the therapeutic window (low drug concentration is difficult to produce a therapeutic effect while the high concentration is easy to produce intolerable toxicity). TDD system based on nanomachines can perform a dual effect that expands the therapeutic window to achieve greater drug efficacy while being less prone to toxicity and reducing the accumulation of drugs in nontarget sites (Bertrand and Leroux 2011; Bae and Park 2011).

Furthermore, drug-loaded nanomachines are gaining increasing attention in the application of overcoming multidrug resistance in cancer therapy. Nanomachines between 10 and 200 nm exhibit prolonged systemic circulation lifetime, sustained drug release kinetics, and better tumor accumulations through both passive and active mechanisms. Those features provide the ability to co-encapsulate multiple therapeutic agents and to synchronize their delivery to the diseased cells (Gao et al. 2012).

Future Directions

Research on TDD systems is still in its infancy. A great deal of efforts is demanded in the areas of nanomachine manufacturing and controlling, drug loading and releasing, and communications-inspired optimal TDD. The existing nanorobotic technologies can hardly satisfy the requirements of controlled directional movement, drug molecules' transportation and releasing, biocompatibility and biodegradability, etc., at the same time. There is still a long way to go before reaching the ultimate goal of clinical applications. Various control methods also have deficiencies in the surveillance and positioning of nanorobots in the human body, especially in the capillary-level microenvironment. As a result, the communications-inspired design of TDD via nanomachines has to make substantial assumptions about the system and channel models for the motion analysis and parameter optimization. These problems will need to be gradually resolved with further experimental verification.

Cross-References

- ▶ [Applications of Molecular Communication Systems](#)
- ▶ [Nanonetworks](#)
- ▶ [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

References

- Artzi L, Bayer EA, Moraïs S (2017) Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat Rev Microbiol* 15(2):83–95
- Bae YH, Park K (2011) Targeted drug delivery to tumors: myths, reality and possibility. *J Con Rel* 153(3):153–198
- Bertrand N, Leroux JC (2011) The journey of a drug carrier in the body: an anatomo-physiological perspective. *J Con Rel* 161(2):152–163
- Chahibi Y, Pierobon M, Akyildiz IF (2015a) Pharmacokinetic modeling and biodistribution estimation through the molecular communication paradigm. *IEEE Trans Biomed Eng* 62(10):2410–2420
- Chahibi Y, Akyildiz IF, Balasubramaniam S, Koucheryavy Y (2015b) Molecular communication modeling of antibody-mediated drug delivery systems. *IEEE Trans Biomed Eng* 62(7):1683–1695
- Chen Y, Zhou Y, Murch R, Kosmas P (2017a) Modeling contrast-imaging-assisted optimal targeted drug delivery: a touchable communication channel estimation and waveform design perspective. *IEEE Trans Nanobioscience* 16(3):203–215
- Chen XZ, Hoop M, Mushtaq F et al (2017b) Recent developments in magnetically driven micro-and nanorobots. *Appl Mater Today* 9:37–48
- Chude-Okonkwo UA, Malekian R, Maharaj BS (2016) Molecular communication model for targeted drug delivery in multiple disease sites with diversely expressed enzymes. *IEEE Trans Nanobioscience* 15(3):230–245
- Femminella M, Realì G, Vasilakos AV (2015) A molecular communications model for drug delivery. *IEEE Trans Nanobioscience* 14(8):935–945
- Feynman RP (1992) There's plenty of room at the bottom. *J Mic Sys* 1(1):60–66
- Gao ZB, Zhang LN, Sun YJ (2012) Nanoparticle-based combination therapy toward overcoming drug resistance in cancer. *Biochem Pharmacol* 83(8):1104–1111
- Guo F, Li P, French JB, Mao Z, Zhao H, Li S, Nama N, Fick JR, Benkovic SJ, Huang TJ (2015) Controlling cell–cell interactions using surface acoustic waves. *Proc Natl Acad Sci* 112(1):43–48
- Hu Q, Katti PS, Gu Z (2014) Enzyme-responsive nanomaterials for controlled drug delivery. *Nanoscale* 6(21):12273–12286

- IEEE NanoCom Standards Development Working Group (2017) IEEE recommended practice for nanoscale and molecular communication framework, IEEE standard 1906.1–2015, 2016
- Kim H, Ali J, Cheang UK, Jeong J, Jin SK, Min JK (2016) Micro manipulation using magnetic microrobots. *J Bionic Eng* 13(4):515–524
- Liu C, Xu T, Xu LP, Zhang X (2017) Controllable swarming and assembly of micro/nanomachines. *Micromachines* 9(1):9–10
- Martel S, Mohammadi M, Felfoul O, Lu Z, Poupponeau P (Apr. 2009) Flagellated magnetotactic bacteria as controlled MRI-trackable propulsion and steering systems for medical nanorobots operating in the human microvasculature. *Int J Robot Res* 28(4):571–582
- Sagnella S, Drummond C (2012) Drug delivery: a nanomedicine approach. *Aust Biochemist* 43:5–8
- Trafton A (2009) Tumors targeted using tiny gold particles. *MIT Tech Talk* 53(24):4–4
- Zhou Y, Chen Y, Murch RD (2018) Simulation framework for touchable communication on NS3Sim. *J Nano Comm Netw* 16:26–36

Dual Connectivity

- [Hyper Cellular Network: Control-Traffic Decoupled Radio Access Network Architecture](#)

Dynamic Pricing

Dusit Niyato
School of Computer Science and Engineering
(SCSE), Nanyang Technological University
(NTU), Singapore, Singapore

Definitions

Dynamic pricing is a mechanism in which the price is determined dynamically and iteratively through dynamic algorithms. The algorithms can yield optimal or equilibrium pricing solutions that maximize the profit of a single monopolistic seller or multiple oligopolistic sellers, respectively.

Economic Modeling and Pricing

Economic and pricing approaches have been applied to address many issues in wireless/radio resource management. In the following, major economic and pricing models used in the wireless literature are presented Luong et al. (2017).

Cost-based pricing: Cost-based pricing is a common pricing strategy to determine the price of wireless resource and service based on calculating the total cost of acquiring the resource and offering the service as well as adding a percentage of the cost as a desired profit. The goal of cost-based pricing is to ensure that the price yields nonnegative profit to the wireless service provider profitable. The total cost generally consists of a fixed cost and a variable cost. The fixed cost is the cost that does not change when the number of sales of the services changes, for example, hardware costs, for example, servers and network devices, and installation. On the contrary, the variable cost incurs depending to the number of product produced or the amount of services used. For example, the resource costs, for example, energy and bandwidth costs. The advantage of the cost-based pricing is the ease of setting the price. The reason is that the price is a function of the internal cost, i.e., the cost required to generate the service Courcoubetis and Weber (2003). However, this pricing strategy does not consider external market factors, for example, the pricing strategies of other providers and the perceive value and willingness to pay of buyers or users.

Differential pricing: The cost-based pricing ignores the requirement and preference of each user. On the other hand, to maximize the profit of providers, differential pricing, also called price discrimination, can be used, utilizing such user's requirement and preference. In the differential pricing, the provider can charge different prices to different users based on their demand and willingness to pay. By setting higher prices for one type of user, the use of the differential pricing explores the user surplus by the provider. However, it can be unfair to cause one type of users to pay a higher price than that of other types of user.

Profit maximization: Profit maximization is the process of determining the output quantity and/or the price that offer the highest profit for a provider Korrapati (2014). The pricing based on profit maximization concept is described in the following. Assume that a service provider needs to determine the amount of radio resource or wireless service units to be sold which is denoted by q and the price p for their users. The profit of the provider is $\pi = R(p, q) - C(q)$, where $R(\cdot, \cdot)$ is the total revenue and $C(\cdot)$ is the total cost. The total cost includes a fixed cost and a variable cost. The revenue is the amount of money that the provider receives from selling q resource units to its users. The optimal quantity of radio resource units, i.e., q^* , is determined such that the profit is maximized, i.e., $q^* = \max_q \pi$. The optimal quantity allows to find the optimal price based on the demand curve. The demand curve is typically a linear curve to show the relationship between the price of a resource unit and the quantity of resource units that users are willing to buy. A general demand curve can be expressed as $p = a - bq$, where a and b are proper parameters. Thus, at $q = q^*$, the optimal price is $p^* = a - bq^*$.

Game Theoretic Pricing

Game theory is the study of multiparticipant decision-making problems in which a choice of a participant, i.e., a player, potentially affects the interests of other participants Gibbons (1992). In the following, game theoretic models are briefly reviewed. Some important terminologies are defined below Shi et al. (2012).

- **Player:** A player is a participant which makes a decision in the game.
- **Payoff:** A payoff, i.e., a utility, a profit, or an interest, reflects the desired outcome of the player.
- **Strategy:** Player's strategy is a set of actions/instructions that the player can follow to achieve the desired outcome. The payoff is a function of not only the player's own action, but also the actions of others.

- **Rationality:** A player is rational if its strategy always aims at maximizing its own payoff.

Noncooperative game: In a noncooperative game, each player maximizes only its own payoff without considering the payoff of the other players nor the social welfare of the networks AlSkaif et al. (2015). In this game, the players are self-interested, and they do not form coalitions or make agreements with each other.

Consider a radio resource market in which providers as the sellers compete for selling radio resource units, for example, spectrum, to users. The sellers are typically selfish. As such, the market consisting of the providers selling the spectrum can be modeled as a non-cooperative game. The providers can adjust prices of the spectrum and therefore the prices are the strategies. Assume that there are N players, and P_i is a set of pricing strategies of player i , where $P = P_1 \times \dots \times P_N$. $\times$ is the Cartesian product of the sets of strategies of each player. Let $p_i \in P_i$ be the pricing strategy of player i . A vector of strategies of N players is $\mathbf{p} = (p_1, \dots, p_N)$, and a vector of corresponding payoffs is $\pi = (\pi_1(\mathbf{p}), \dots, \pi_N(\mathbf{p}))$, where $\pi_i(\mathbf{p})$ is the payoff of player i given the player's chosen strategy and strategies of all the other players. Each player optimizes its best strategy p_i^* which maximizes its payoff. A set of strategies $\mathbf{p}^* = (p_1^*, \dots, p_N^*) \in P$ is the Nash equilibrium of the noncooperative game if no player can gain higher payoff by changing its own strategy when the strategies of the others remain the same Friedman (1988), i.e.,

$$\forall i, p_i \in P_i : \pi_i(p_i^*, \bar{\mathbf{p}}_i^*) \geq \pi_i(p_i, \bar{\mathbf{p}}_i^*), \quad (1)$$

where $\bar{\mathbf{p}}_i = (p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_N)$ is a vector of prices of all players except player i .

Stackelberg game: The noncooperative game assumes that players chooses and releases their pricing strategies simultaneously. Also, all the players know each other's strategies at the same time. However, this may not always hold in real markets. Therefore, sequential games can be used

in which players can choose and release their strategies following a certain predefined order. This situation can be modeled as a Stackelberg game Amir and Grilo (1999). In the Stackelberg game, the players are divided into leaders and followers. The follower players decide its own strategic choices after observing the strategies of leader players Kim (2014). It was proved that even if the leader players can choose their strategies first, their payoffs are not less than those at the Nash equilibrium Han (2012), i.e., because of the first-mover advantage. The following provides the definition and properties of the Stackelberg game. Assume that there are two radio resource sellers 1 and 2 in the market. P_1 and P_2 are the sets of pricing strategies of sellers 1 and 2, respectively. Seller 1 chooses its pricing strategy p_1 from set P_1 to maximize its payoff or profit function $\pi_1(p_1, p_2)$, and seller 2 chooses its pricing strategy p_2 from set P_2 to maximize its payoff function $\pi_2(p_1, p_2)$. Also, assume that seller 2 selects its strategy before seller 1 decides its selection, and hence seller 1 is able to observe the strategy of seller 2. Seller 2 is then the leader, and seller 1 is the follower. We have the following definition Leitmann (2013):

Definition 1

If there exists a mapping $F : P_2 \rightarrow P_1$ such that, for any fixed $p_2 \in P_2$, $\pi_1(F(p_2), p_2) \geq \pi_1(p_1, p_2)$, $\forall p_1 \in P_1$, and if there exists $p_{2s2} \in P_2$ such that $\pi_2(F(p_{2s2}), p_{2s2}) \geq \pi_2(F(p_2), p_2)$, then the pair $(p_{1s2}, p_{2s2}) \in P_1 \times P_2$, where $p_{1s2} = F(p_{2s2})$, is called a Stackelberg strategy pair.

Definition 1 presents that the Stackelberg strategy is optimal for the leader when the follower responds to the leader with the follower's optimal strategy.

Case Study of Dynamic Spectrum Pricing

Pricing Models

Market-Equilibrium-Based Pricing: In the market-equilibrium pricing model, the primary service is not aware of other primary services.

Thus, at the seller side, the primary service naively sets the price according to the spectrum demand of the secondary service. This price setting is based on the willingness of the primary service to sell spectrum which is generally determined by the *supply function*. For a given price, supply function indicates the size of radio spectrum to be shared by a primary service with the secondary service. At the buyer side, the willingness of a secondary service to buy spectrum is determined by the *demand function*. Again, for a given price, demand function determines the size of radio spectrum required by a secondary service. In this spectrum trading, market-equilibrium price denotes the price for which spectrum supplied by the primary service is equal to the spectrum demand from the secondary service. This market-equilibrium price ensures that there is no excess supply in the market and spectrum supply meets all spectrum demand.

Competitive Pricing: In the competitive pricing model, a primary service is aware of the existence of other primary services and all of the primary services compete with each other to achieve the highest individual profit. The competition here occurs in terms of spectrum pricing. In particular, given the spectrum prices offered by other primary services, one primary service chooses the price for its own spectrum so that its individual profit is maximized. This competitive spectrum pricing can be considered as a price war in an oligopoly market. In this market structure, the firms choose their price noncooperatively. The decision of each seller is influenced by other sellers' actions and action of one seller may be observed by other sellers. A general description for this oligopoly market is as follows. There are N sellers, i.e., primary services, in the market offering the same or differentiated product, i.e., spectrum. The seller competes with each other by adjusting the price of product selling to the buyer, i.e., secondary service. In this competitive situation, when one seller varies the price, demand for that product will change and other sellers will observe and adjust their prices to gain higher profit. This competition in oligopoly market can be modeled by noncooperative game.

Cooperative Pricing: In the cooperative pricing model, it is assumed that all of the primary services know each other and they fully cooperate (i.e., collude) to obtain the highest total profit by selling spectrum to the secondary service. In an actual environment, to achieve this full cooperation, extensive communication would be required among all primary services. The demand function for spectrum F_i at the secondary service can be obtained using $\frac{\partial U(\mathbf{b})}{\partial b_i} = 0$ as:

$$D_i(\mathbf{p}) = \frac{(k_i^{(s)} - p_i)(v(N-2)+1) - v \sum_{i \neq j} (k_j^{(s)} - p_j)}{(1-v)(v(N-1)+1)},$$

where $\mathbf{p}$ denotes a vector of prices offered by all primary services in the market (i.e., $\mathbf{p} = [p_1 \dots p_i \dots p_N]^T$).

Revenue and Cost Functions for a Primary Service: For a primary service, there are two sources of revenue – from primary users and secondary users. However, a cost is involved which is a function of QoS performance degradation of ongoing primary users due to sharing the radio spectrum with secondary service. We assume that the primary users are charged at a flat rate for a guaranteed amount of bandwidth. However, if the required bandwidth cannot be provided, a primary service offers “discount” to the users, and this is considered as the cost of sharing spectrum with the secondary service. Let R_i^l denote the revenue gained from primary users served by primary service i , R_i^s denote the revenue gained from sharing spectrum with secondary users, and C_i denote the cost due to QoS degradation of primary users. Then, the revenue and cost functions can be defined as: $R_i^s = p_i b_i$, $R_i^l = c_1 M_i$, $C_i(b_i) = c_2 M_i \left(B_i^{req} - k_i^{(p)} \frac{W_i - b_i}{M_i} \right)^2$, where b_i and p_i denote, respectively, the spectrum size shared with secondary service and the corresponding price, and c_1 and c_2 denote constant weights for the revenue and cost functions at the primary service, respectively. Here, B_i^{req} denotes bandwidth requirement per user, W_i denotes spectrum size, M_i denotes the number of ongoing primary users, and $k_i^{(p)}$ denotes spectral efficiency of wireless transmission for primary service i . Note that revenue from the primary users is a linear function of the number of

concurrent users, while revenue from secondary users is a linear function of the shared spectrum size given the spectrum price. The cost is proportional to the square of the difference between bandwidth requirement and allocated bandwidth to a primary user.

Solution of Pricing Models

Solution of Market-Equilibrium Pricing Model:

For each primary service, the spectrum supply function can be derived based on a profit maximization problem. The solution of this optimization formulation is the optimal spectrum size b_i to be shared with the secondary service for a given price p_i . Based on revenue (i.e., R_i^l and R_i^s) and cost (i.e., C_i), profit P_i of a particular primary service i owning spectrum F_i can be expressed as: $P_i = p_i b_i + c_1 M_i - c_2 M_i \left(B_i^{req} - k_i^{(p)} \frac{W_i - b_i}{M_i} \right)^2$. To obtain the optimal spectrum size to be shared, we differentiate the profit function with respect to b_i (when p_i is given) and obtain that $\frac{\partial P_i}{\partial b_i} = 0 = -p_i + 2c_2 M_i \left(B_i^{req} - k_i^{(p)} \frac{W_i - b_i}{M_i} \right) \frac{k_i^{(p)}}{M_i}$.

Spectrum supply is given by the optimal value of b_i^* which is a function of price p_i . The supply function can be expressed as: $S_i(p_i) = W_i - \frac{M_i}{k_i^{(p)}} \left(B_i^{req} - \frac{p_i}{2c_2 k_i^{(p)}} \right)$. The market-equilibrium (i.e., solution) is defined as the price p_i^* at which spectrum supply equals spectrum demand, i.e., $S_i(p_i^*) = D_i(\mathbf{p}^*)$, $\forall i$, where the vector $\mathbf{p}^* = [\dots p_i^* \dots]^T$ denotes the market-equilibrium prices for all primary services.

Solution of Competitive Pricing Model: We use a noncooperative game to model the price competition among primary services. The players (i.e., sellers in an oligopoly market) in this game are the primary services. The strategy of each of the players is the price per unit of spectrum. The payoff for each primary service i (denoted by P_i) is the individual profit due to selling spectrum to the secondary service.

Again, based on the demand, revenue, and cost functions, the individual profit of each primary service can be expressed as follows: $P_i(\mathbf{p}) =$

$R_i^s + R_i^l - C_i$, where $\mathbf{p}$ denotes a vector of prices offered by all of the players (i.e., primary services) in the game. We consider the Nash equilibrium as a solution of this price competition. In this case, the Nash equilibrium is obtained by using the best response function which is the best strategy of one player given others' strategies. The best response function of primary service i , given a vector of prices offered by other primary services $\mathbf{p}_{-i}$, is defined as: $B_i(\mathbf{p}_{-i}) = \arg \max_{p_i} P_i(p_i, \mathbf{p}_{-i})$.

The vector $\mathbf{p}^* = [\dots p_i^* \dots]^T$ denotes a Nash equilibrium (i.e., solution) of this game on competitive pricing if and only if $p_i^* = B_i(\mathbf{p}_{-i}^*) \forall i$, where $\mathbf{p}_{-i}^*$ denotes the vector of best responses for player j for $j \neq i$. Mathematically, to obtain the Nash equilibrium, we have to solve the following set of equations: $\frac{\partial P_i(\mathbf{p})}{\partial p_i} = 0$ for all i . In this case, the size of the shared bandwidth b_i in the individual profit function is replaced with spectrum demand $D_i(\mathbf{p})$, and then the profit function can be expressed as: $P_i(\mathbf{p}) = p_i D_i(\mathbf{p}) + c_1 M_i - c_2 M_i \left(B_i^{req} - k_i^{(p)} \frac{W_i - D_i(\mathbf{p})}{M_i} \right)^2$.

Then, using $\frac{\partial P_i(\mathbf{p})}{\partial p_i} = 0$, we obtain that $2c_2 k_i^{(p)} D_2 \left(B_i^{req} - k_i^{(p)} \frac{W_i - (D_1(\mathbf{p}_{-i}) - D_2 p_i)}{M_i} \right) + D_1(\mathbf{p}_{-i}) - 2D_2 p_i = 0$. Recall that the demand function can be expressed as $D_1(\mathbf{p}) = D_1(\mathbf{p}_{-i}) - D_2 p_i$. The solution p_i^* , which is a Nash equilibrium, can be obtained by solving the above set of linear equations by using a numerical method when all the above parameters are available. Then, given a vector of prices $\mathbf{p}^*$ at the Nash equilibrium, the size of the shared spectrum can be obtained from the spectrum demand function $D_i(\mathbf{p}^*)$.

Solution of Cooperative Pricing Model: For the cooperative pricing model, an optimization problem is formulated to obtain the optimal price which provides the highest total profit for all primary services. This optimization problem can be expressed as follows:

$$\text{Maximize : } \sum_{i=1}^N P_i(\mathbf{p}), \quad (2)$$

$$\text{Subject to : } W_i \geq b_i \geq 0, \quad (3)$$

$$p_i \geq 0, \quad (4)$$

where the total profit for all the primary services is given by $\sum_{i=1}^N P_i(\mathbf{p})$. Since the constraint in Eq. (3) can be written as $W_i \geq D(\mathbf{p}) \geq 0$, the Lagrangian can be expressed as: $L(\mathbf{p}) = \sum_{i=1}^N P_i(\mathbf{p}) - \sum_{j=1}^N \lambda_j (-p_j) - \sum_{k=1}^N \mu_k (D_k(\mathbf{p}) - W_k) - \sum_{l=1}^N \sigma_l (-D_l(\mathbf{p}))$, where λ_j , μ_k , and σ_l are Lagrange multipliers for the constraints in Eqs. (3) and (4), respectively. Using Kuhn-Tucker conditions, we can obtain the vector of optimal prices $\mathbf{p}^*$ such that the total profit of all the primary services is maximized.

Dynamic Pricing Algorithms

In a practical cognitive radio environment, a primary service may not have the complete network information. Therefore, the primary service must learn the behavior of other entities from the history, and a distributed price adjustment algorithm is required which would gradually reach the solution.

Distributed Implementation of Market-Equilibrium Dynamic Pricing: The solution of this pricing model is given by market-equilibrium which is defined as the price at which spectrum supply from a primary service is equal to spectrum demand from the secondary service. Therefore, the price offered by each primary service is gradually adjusted in a direction that minimizes the difference between spectrum demand and supply. This process works as follows. The spectrum price is initialized to $p_i[0]$ and this price is sent to the secondary service. The secondary service replies with the size of spectrum demand which is computed from the demand function $D_i(\mathbf{p}[t])$ for spectrum F_i . Then, the primary service computes the size of the supplied spectrum $S_i(p_i[t])$. To obtain the price in the next iteration, the difference between spectrum demand and supply at time t is computed, weighted by learning rate α_i , and added to the price in the current iteration. This process repeats until the difference of prices in current iteration t

and next iteration $t + 1$ becomes less than the threshold ε (e.g., $\varepsilon = 10^{-5}$). This price adjustment in each iteration can be expressed as: $p_i[t + 1] = p_i[t] + \alpha_i(D_i(\mathbf{p}[t]) - S_i(p_i[t]))$, where $\mathbf{p}[t]$ is the vector of prices at iteration t , i.e., $\mathbf{p}[t] = [p_1[t] \cdots p_i[t] \cdots p_N[t]]^T$.

Distributed Implementation of Competitive Dynamic Pricing: The solution of this dynamic pricing model is given by the Nash equilibrium. It is defined as the prices which satisfies all of the primary services. We assume that a primary service cannot observe the prices of other primary services. Therefore, each primary service can use only local information and spectrum demand information from the secondary service to adjust its strategy. This distributed competitive dynamic pricing works as follows. The spectrum price is initialized to $p_i[0]$ and this price is sent to the secondary service. The secondary service replies with the size of spectrum demand. Also, the primary service estimates marginal individual profit and uses this together with spectrum demand information to compute the spectrum price in the next iteration. The relationship between the prices in the current and the next iterations can be expressed as: $p_i[t + 1] = p_i[t] + \alpha_i \left(\frac{\partial P_i(\mathbf{p}[t])}{\partial p_i[t]} \right)$, where α_i is the learning rate. Let $\mathbf{p}_{-i}[t]$ denote the vector of prices of all primary services except service i at iteration t . To estimate the marginal individual profit, a primary service can observe the marginal spectrum demand for a small variation in price ξ (e.g., $\xi = 10^{-4}$). That is,

$$\frac{\partial P_i(\mathbf{p}[t])}{\partial p_i[t]} \approx \frac{P_i([\cdots p_i[t] + \xi \cdots]) - P_i([\cdots p_i[t] - \xi \cdots])}{2\xi} \quad (5)$$

Note that due to this estimation of marginal individual profit, communication overhead between a primary service and the secondary service is larger than that in case of market-equilibrium dynamic pricing model. Nonetheless, communication among primary services is not required for distributed competitive dynamic pricing algorithm.

Distributed Implementation of Cooperative Dynamic Pricing: The solution of this dynamic pricing model yields an optimal price for which the total profit of all primary services is maximized. Again, we assume that a primary service can observe the variation of spectrum demand from the secondary service. In addition, to achieve the highest total profit, primary services can exchange information on current profit among each other. This is possible only if all the primary services are fully cooperative. The distributed cooperative dynamic pricing then works as follows. The spectrum price is initialized to $p_i[0]$ and then it is sent to the secondary service. The secondary service replies with the spectrum demand. Then, the primary service estimates marginal total profit by exchanging information with the other primary services and uses this together with the spectrum demand from secondary service to compute spectrum price in the next iteration. The price update is done as: $p_i[t + 1] = p_i[t] + \alpha_i \left(\frac{\partial \sum_{j=1}^N P_j(\mathbf{p}[t])}{\partial p_i[t]} \right)$. Again, similar to that in Eq. (5), to estimate marginal total profit, a primary service can observe the marginal total profit through information exchange among all primary services for a small variation in price ξ as:

$$\frac{\partial \sum_{j=1}^N P_j(\mathbf{p}[t])}{\partial p_i[t]} \approx \frac{\sum_{j=1}^N P_j([\cdots p_i[t] + \xi \cdots]) - \sum_{j=1}^N P_j([\cdots p_i[t] - \xi \cdots])}{2\xi}$$

Compared with the market-equilibrium and competitive dynamic pricing models, this estimation of marginal total profit incurs the largest

communication overhead since the profit information needs to be exchanged among all the primary services.

References

- AlSkaif T, Zapata MG, Bellalta B (2015) Game theory for energy efficiency in wireless sensor networks: latest trends. *J Netw Comput Appl* 54(1):33–61
- Amir R, Grilo I (1999) Stackelberg versus cournot equilibrium. *Games Econom Behav* 26(1):1–21
- Courcoubetis C, Weber R (2003) Cost-based pricing. Wiley, pp 161–194. <https://doi.org/10.1002/0470867175.ch7>
- Friedman JW (1988) A non-cooperative equilibrium for supergames. Cambridge University Press
- Gibbons R (1992) A primer in game theory. Harvester Wheatsheaf
- Han Z (2012) Game theory in wireless and communication networks: theory, models, and applications. Cambridge University Press
- Kim S (2014) Game theory applications in network design. IGI Global
- Korrapati R (2014) Validated management practices. A.H.W. Sameer series. Diamond Pocket Books <https://books.google.com.sg/books?id=z4psBQAAQBAJ>
- Leitmann G (2013) Multicriteria decision making and differential games. Springer
- Luong NC, Wang P, Niyato D, Wen Y, Han Z (2017) Resource management in cloud networking using economic analysis and pricing models: a survey. *IEEE Commun Surv Tutor* 19(2):954–1001. <https://doi.org/10.1109/COMST.2017.2647981>
- Shi HY, Wang WL, Kwok NM, SY C (2012) Game theory for wireless sensor networks: a survey. *Sensors* 12(7):9055–9097

Dynamic Spectrum Access

- [Database-Based Spectrum Access](#)
- [Matching for Cooperative Spectrum Sharing](#)

Dynamic Spectrum Access in Multiple Primary Networks

Ye Wang and Yi Yang
Harbin Institute of Technology (Shenzhen),
Shenzhen, China

Cognitive radio technology is an effective way to solve the contradiction between the rapid devel-

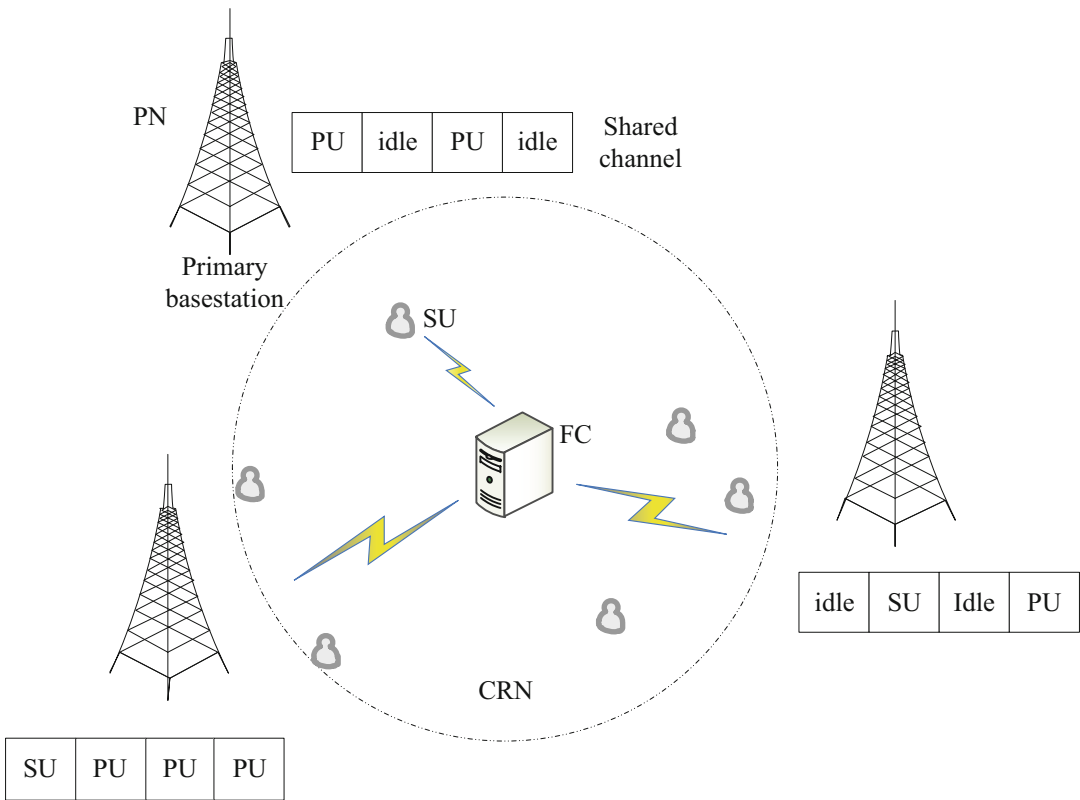
opment of wireless applications and inefficient spectrum allocation. However, the preemptive priority of the primary users (PUs) and the space time variation characteristics of the idle channel lead to the unstable performance of the cognitive radio network (CRN). With the development of network convergence technology and the popularity of heterogeneous networks, the cognitive radio network with multiple primary networks (CRNMPNS) is becoming possible. CRNMPNS provides more spectrum resources and network selections.

Network Topology

Multiple primary cognitive radio networks consist of PRNs (primary radio networks) and CRNs (cognitive radio networks). Its topological structure is shown in Fig. 1. Some PRNs are distributed in a certain geographical area, and CRNs are located in the area around them.

CRN consists of SUs and Fusion Center (FC). SUs are divided into in-band SUs and out-of-band SUs. In-band SUs refer to SUs that have access to authorized channels, and those that have not access are called out-of-band SUs. For in-band SUs, it is not necessary to participate in spectrum sensing all the time. It is only necessary to periodically sense whether the currently occupied channel has a PU's arrival while maintaining normal communication. If a PU arrives, it immediately exits the current channel and sends access request to FC again. For out-of-band SUs, FC allocates spectrum sensing. If there is an idle channel, FC allocates the right to access the authorized channel. Because FC has known the distribution of in-band SUs, the sensing channel only includes the channel occupied by PU and idle channel.

The workflow diagram of multiple primary cognitive radio networks is shown in Fig. 2. Firstly, according to the current network state, FC adopts a certain cooperative sensing scheduling scheme. According to the sensing scheduling, SUs sense the corresponding PRNs. Then, SUs are sensed separately by single-point sensing, and the sensing results are fused by FC according to



Dynamic Spectrum Access in Multiple Primary Networks, Fig. 1 Topological structure of CRNMPNS

certain fusion criteria to determine the current final channel state of PRNs. When SU initiates an access request, FC chooses the best PRN for dynamic spectrum access according to the corresponding network selection scheme. Finally, the performance of the CRN network is evaluated, such as PU collision probability, CRN throughput, SU congestion rate, SU dropping rate, PU congestion rate, and other performance indicators. Network Selection of Multiple primary Networks is a key technology in dynamic spectrum access.

Network Selection

Network selection is divided into offline network selection and online network selection.

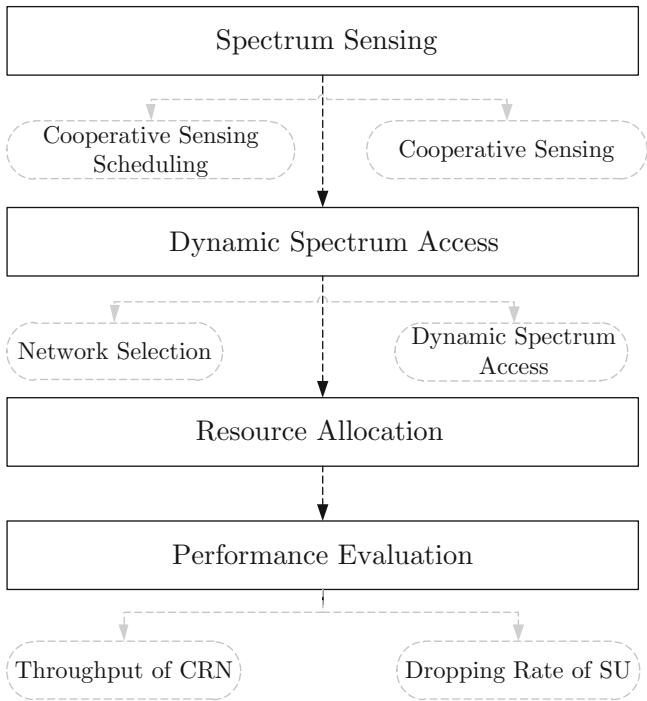
Offline Network Selection

The problem of off-line network selection is that in the current network state, each SU has equal probability of accessing PRNs. Therefore, the performance of CRN system varies greatly with network selection. The network selection problem in multiauthorized cognitive radio networks can be described as shown in Fig. 3.

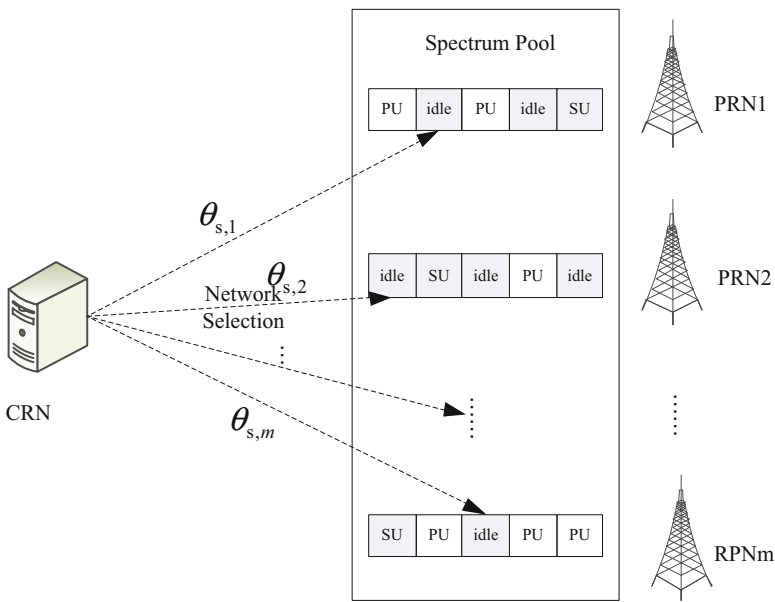
Offline network selection scheme mainly includes random network selection scheme, greedy network selection scheme based on number of idle channels, and weighted network selection scheme.

Random network selection scheme refers to the random and equal probability selection of an authorized network access from multiple primary networks. Random network selection scheme is the simplest, and almost no SU control. However, due to randomness, the differences among PRNs are not taken into account, such as the number

Dynamic Spectrum Access in Multiple Primary Networks, Fig. 2
work flow chart of CRNMPNS



Dynamic Spectrum Access in Multiple Primary Networks, Fig. 3
Network selection description diagram



of shared channels, the constraints of PRNs on CRNs.

The greedy network selection scheme based on the number of idle channels is that the SU to be accessed chooses the authorized network access

scheme with the largest number of idle channels at the current time. Greedy network selection scheme always chooses the best PRN access in the current network, which can achieve load balancing effect to a certain extent and reduce SU

dropping rate. Under unconstrained conditions, it should be the optimal network selection scheme.

The weighted network selection scheme is based on the weighted PRNs. For the PRNs with more shared channels, the weighted PRNs are larger, and access is based on the proportion of the weighted PRNs. Weighted network selection scheme can effectively improve spectrum utilization, reduce SU dropping rate, and improve SU quality of service.

Online Network Selection

Because of the dynamic changes of network environment, when some factors change, the off-line network selection scheme will not meet the needs of dynamic environment. Therefore, online network selection is proposed, that is, network selection scheme and simulation environment will be dynamically adjusted with the change of parameters.

Because the problem of off-line network selection is based on the prior conditions of PU and SU, the corresponding expression of measurement index is obtained by theoretical deduction. Once the network environment changes, it needs to be recalculated and deduced to obtain closed solution, so the network cannot be dynamically adjusted. Moreover, the priori conditions of PU and SU such as arrival rate and service time are unknown in real environment. Online network selection do not need the priori conditions of PU and SU that makes this network selection scheme fits better in dynamic environment.

Online network selection scheme mainly includes ONUP network selection scheme, ONUP & Greedy network selection, and ONUP & Min network selection.

ONUP network selection scheme is to update the network selection scheme online. The basic idea of this scheme is to update the collision probability of PU as long as the corresponding network state changes. Because the impact factors of the collision probability of PU include the number of PU access channels and the number of PU preempted channels from SU, when PU access

the idle channel or PU preempt SU channels, the collision probability will be recalculated.

ONUP & Greedy network selection scheme adds a judgment process on the basis of ONUP network selection. When the collision probability is larger than the upper limit, judgment is added to find out the largest number of idle channels in the current state within a certain range of guaranteeing the maximum collision probability. If the number is greater than 2, the priority is 1 in the current state, and if it is not satisfied, the priority is infinitesimal. When the number of idle channels is more than 2, it is guaranteed that the transmission will not be interrupted because of the random arrival of PU when the idle channel is accessed.

ONUP & Min network selection scheme chooses the access with the lowest collision probability when the collision probability is greater than the collision probability.

Summary

In this chapter, we introduced the application background of multiple primary networks. The topological structure of multiple primary networks is based on heterogeneous networks. Unlike other networks, dynamic spectrum access of multiple primary networks has network selection process. Two kinds of network selection scheme offline and online selection are introduced. Online selection fits better in dynamic wireless communication environment than offline selection.

Cross-References

- [Database-Based Spectrum Access](#)
- [Dynamic Spectrum Access Through Cognitive Radio Networking](#)

Dynamic Spectrum Access Through Cognitive Radio Networking

Zhipeng Wang and Bin Cao
EIE School, Harbin Institute of Technology,
Shenzhen, China

Synonyms

Cognitive radio; Dynamic spectrum management

Definitions

Dynamic spectrum access is a radio technology that can dynamically select a licensed spectrum within a proper access right to communicate in a time, space, and frequency domain. This is a communication mechanism that allows unauthorized user (also known as Secondary user, SU) to access the licensed spectrum for communication without interfering with the authorized user (Primary user, PU) communication.

Historical Background

Due to the limited spectrum and strict government control, the spectrum resources are scarce and expensive. Dynamic spectrum access is a powerful technology to support the development of wireless communications and network. The concept of dynamic spectrum access was first proposed by the NeXt Generation Program, which was established in 2003 by Defense Advanced Research Projects Agency (Akyildiz Ian et al. 2006).

Later on, the IEEE Standard Association (IEEE SA), IEEE Communications Society (ComSoc), and IEEE Electromagnetic Compatibility (EMC) jointly established the IEEE P1900 Standardization Committee and formally defined the concept of dynamic spectrum access in 2005. The IEEE P1900 goal is to support the standardization of next-generation

radio and advanced spectrum management technologies. In 2007, the IEEE Standards Board approved IEEE P1900 reorganization as Standards Coordinating Committee 41 (SCC41), Dynamic Spectrum Access Networks (DySPAN). Then, SCC41 was directly under the IEEE ComSoc Standards Board and renamed the IEEE Dynamic Spectrum Access Networks Standards Committee (DySPAN-SC) in 2010. IEEE DySPAN-SC redefines the scope of SCC41 research (Prasad et al. 2008), including dynamic spectrum access radio systems and networks with the focus on improved spectral efficiency; new technology of dynamic spectrum access based on spectrum interference management; network management and information sharing between multiple network technologies.

In 2010, Federal Communications Commission (FCC) issued a policy to open digital TV white space (TVWS) to unauthorized commercial device, allowing unlicensed users to use TVWS with authorized users, without disrupting digital TV broadcasts and wireless microphone users (FCC 2010). On June 13, 2012, the FCC released Notice of Proposed Rulemaking (NPRM) on the 4.9GHz public safety radio. Based on the work of the FCC, IEEE 802.11 has established the IEEE 802.11af working group to standardize WLAN services for the TVWS band.

In addition, the “Dynamic Radio for IP Services in Vehicular Environments, DRiVE” project by the European Telecommunications Standards Institute (ETSI) aims to use the public coordination channel to achieve dynamic sharing of spectrum resources between heterogeneous networks (Khattab et al. 2013). As a preliminary research on the key technologies of dynamic spectrum sharing, this project provides a sufficient technical reserve for dynamic spectrum access. “Specific Efficient Uni- and Multicast Over Dynamic Radio Networks for IP services in Vehicular Environments, OverDRiVE” project in ETSI aims to enhance and coordinate the Universal Mobile Telecommunication System (UMTS) wireless network to ensure efficient use of spectrum in vehicle multimedia services. As an extension of the DRiVE/OverDRiVE project, ETSI launched the “End to End

Reconfigurability, E2R” research project, which aims to use software defined radio to fuse different types of wireless networks in the future, enabling end-to-end reconfigurability and coexistence of multiple wireless access systems such as cellular, wireless LAN, and digital video broadcasting.

Foundations

Cognitive Radio

Based on software radio technology, Joseph Mitola 2000 was the first to give the following definition of cognitive radio in his PhD thesis (Mitola 2000). In his paper, he explained how cognitive radio could improve wireless services through Radio Knowledge Representation Language (RKRL). Cognitive radio should make full use of the intelligent computing capabilities of wireless digital devices and related networks in radio resources and communications to detect user communication demands and provide the most appropriate radio resources and wireless services based on these needs. As shown in Fig. 1, in a cognitive radio system, the system constantly observes changes in the environment and state and continuously learns to formulate corresponding optimal strategies to process the events. Therefore, in order to achieve dynamic

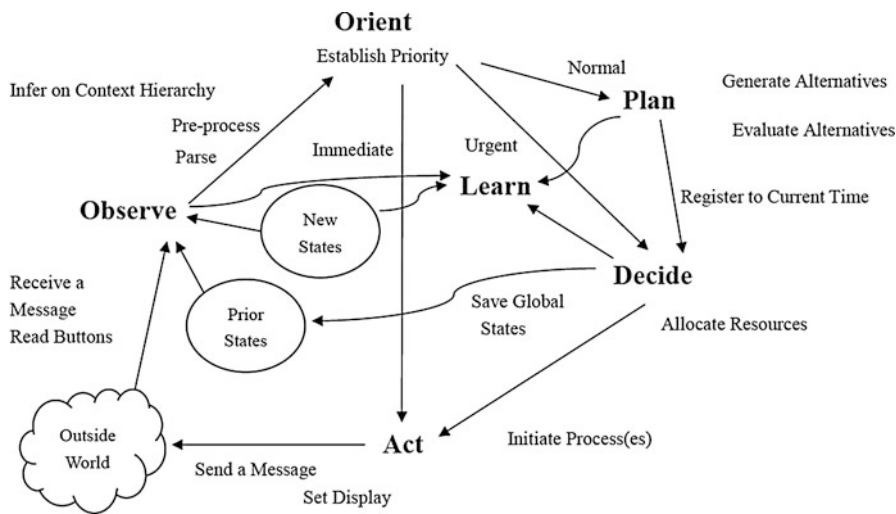
spectrum access, the most basic condition is that the SU is equipped with cognitive radio. Those SU form a cognitive radio network (CRN). In this context, these three paradigms are known as interweave CRN, underlay CRN, and overlay CRN (Goldsmith et al. 2009).

Interweave CRN

Interweave CRN differs from the others CRN in that PUs has absolute priority in using spectrum, which means that when the PU is using this licensed spectrum, SUs cannot access the band. Therefore, the Interweave CRN is also known as opportunistic spectrum access, in which case the SU can exploit spectrum gaps in time, space, or frequency domain. As shown in Fig. 2, SUs use cognitive radio technology to perceive the surrounding spectral environment and then select one or more idle spectrum. The SUs transmitters convert them into a selected spectrum for transmission. Due to opportunistic spectrum access, SUs are highly unstable in communication, and SUs transmission performance degrades when PU reclaim the spectrum.

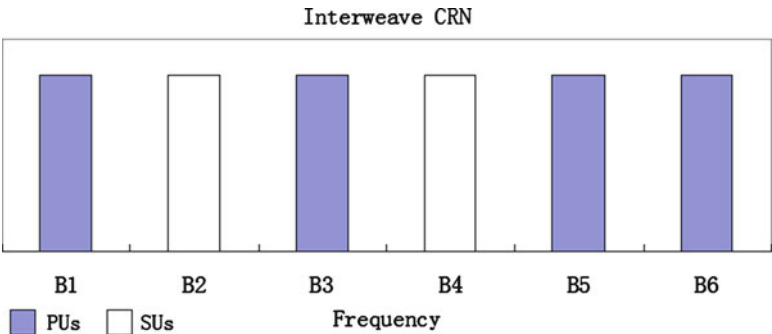
Underlay CRN

The underlay CRN allows SUs to access licensed spectrum regardless of whether the PUs are accessing, and the PUs can tolerate interference from all SUs, which means some threshold

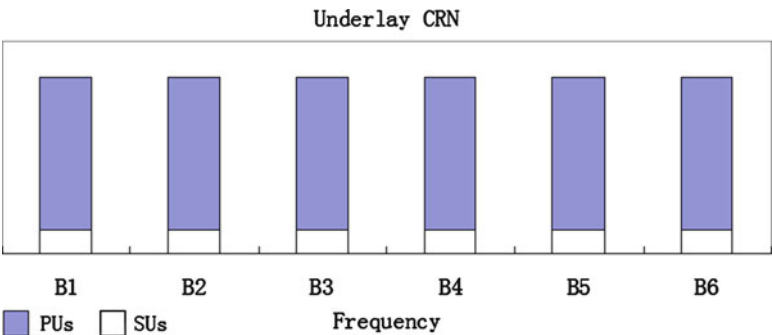


Dynamic Spectrum Access Through Cognitive Radio Networking, Fig. 1 Cognition cycle

Dynamic Spectrum Access Through Cognitive Radio Networking, Fig. 2
Interweave CRN spectrum



Dynamic Spectrum Access Through Cognitive Radio Networking, Fig. 3
Underlay CRN spectrum by using UWB

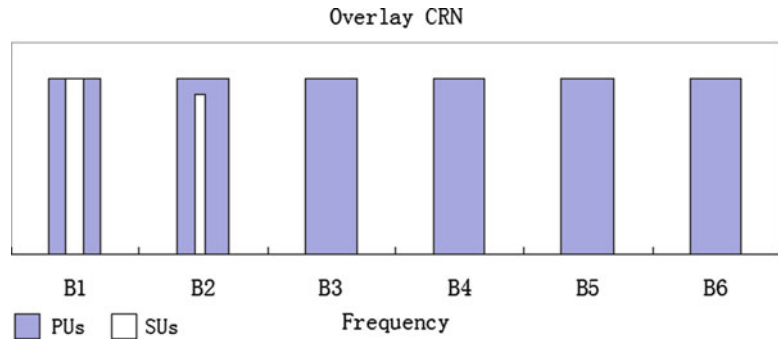


constraints are required in PUs receivers. There are two methods to satisfy this constraint. In the first method, ultra-wide band (UWB) technology that allows signals to have a bandwidth of GHz by directly modulating the impulse pulses with very steep rise and fall times is used to enable SUs transmit information to propagate over a wide spectrum. As shown in Fig. 3, it causes the interference of the SUs on the licensed spectrum far below the threshold in the PU receiver. This method is mainly used for short-range communication due to the low transmission power of SUs. The second method is called the interference temperature – interference temperature that a model quantifies and manages interference recommended by the FCC at the end of 2003. In the interference temperature mechanism, the interference temperature is used to characterize the sum of the interference generated by the SUs and the system noise power at the PUs receivers. The SUs can use the licensed spectrum if the interference at the PU receiver does not exceed the interference temperature threshold. In this way, SU can transmit data at a higher power on the licensed spectrum.

Overlay CRN

Compared with the underlay CRN, the overlay CRN also allows SUs to coexist with the PUs. However, the difference between overlay CRN and underlay CRN is that SUs can use the information of the PUs to select and optimize the corresponding communication methods. In the underlay CRN and interweave CRN, the PUs do not need to know whether the SUs exist or not, and there is no interaction between them. In the overlay CRN, PUs can gain benefits by establishing a cooperative relationship with the SUs, and the SUs can use the PUs spectrum for transmission. The cost of the SUs using the licensed spectrum to obtain transmission opportunities is to provide services to the PUs so that the PUs can transmit its information to its intended destination with greater reliability. In this way, PUs and SUs mutually benefit from cooperation, i.e., the PUs and the SUs cooperate to achieve mutual benefit. The negotiation of SUs and PUs is required to implement the overlay CRN. As shown in the Fig. 4, the PUs enable SUs to use the spectrum by completely or sharing a spectrum. When the SU and the PU share a part

**Dynamic Spectrum
Access Through
Cognitive Radio
Networking, Fig. 4**
Overlay CRN spectrum



D

of the spectrum, channel coding or network coding is adopted to eliminate interference between them.

Spectrum Sensing

Different from the traditional wireless communication system, the goal of dynamic spectrum access of CRN is to improve the utilization of spectrum resources and satisfy the communication of SUs under the premise of satisfying the interference requirements of the PUs. In order to meet the requirements in the CRN, the SUs must have more accurate spectrum sensing capability than traditional wireless communication users. Spectrum sensing can generally be divided into local spectrum sensing and cooperative spectrum sensing (Ramani and Sharma 2017). The most effective method to improve accurate spectrum sensing at present is cooperative spectrum sensing. Under local spectrum sensing condition (such as energy detection, matched filter, cyclostationary detection, and wavelet detection), the effects of the SNR will deteriorate and limit the sensing in a low signal-to-noise ratio (SNR) environment. In addition, a hidden terminal problem may be caused, that is, SUs are shadowed in severe multipath fading environments or inside buildings with high penetration loss. In these cases, the SUs may not be able to detect the presence of the PUs and will then access the licensed spectrum and cause severe co-channel interference to the PUs. Cooperative spectrum sensing can obtain better sensing accuracy than local spectrum sensing, reduce local spectrum

sensing complexity and detection time, and solve Hidden Stations problem.

According to the difference of the location of the spectrum sensing decision, it is divided into distributed and centralized cooperative spectrum sensing. SUs in different locations simultaneously perform local spectrum sensing and upload the sensing result or sensing data to the fusion center of the upper level. The fusion center then determines the existence of the PUs through some data fusion algorithm, such as hard decision fusion, soft decision fusion, and quantized decision fusion. In distributed cooperative sensing, neighboring users exchange their respective local spectrum sensing results and then make their own judgments and decisions. This approach does not require a central control node and reduces signaling interactions.

Cooperative CRN

Although cooperative spectrum sensing between SUs has satisfactory performance, they have a highly complex implementation. In addition, even if the SUs have accurate sensing capability, the spectrum sensing-based method cannot fully guarantee the transmission performance of the SUs in interweave CRN. Because the SUs must abandon its transmission when they detect the presence of the PU, and sense another spectral hole for transmission. If no spectrum is available, the SUs must stop transmitting. The main reason for this problem is due to the lack of cooperation between PUs and SUs.

SUs can achieve a reasonable cooperative relationship with PUs by providing benefits to PUs,

and SUs obtain the right to authorized spectrum access from PUs. This is called cooperative CRN (CCRN) (Cao et al. 2016). There are two promising approaches in CCRN, cooperative communications and spectrum leasing between PUs and SUs. For cooperative communication in CCRN, the cost of the SUs is that the SUs improve the communication quality of PUs, which one or more SUs can act as relaying nodes for a PU. In this case, the PUs communication links can be established with the help of SUs multihop transmission, which can obtain diversity gain and enhance reception by appropriately combining signals from the multipaths for PUs. As a result, spectral efficiency and utilization are significantly improved. Due to the different channel conditions between the SUs and the PUs, the PU needs to select the SUs that provide the appropriate relay service. The SU uses the additional transmit power to forward the PU information, so that the SU needs to evaluate its gain and cost and select the most suitable PU as its partner.

Since the PUs can transmit by purchasing an authorized spectrum, the SU can lease the spectrum from the PU to achieve the communication requirements. Similarly, the incentive for PU in CCRN comes from the rental spectrum to obtain revenue. In a spectrum-leasing model, the biggest challenge is the pricing issue, which is due to the competition between the PUs selling spectrum and the auction spectrum between SUs. This situation often leads to very complicated pricing strategies. In order to solve the pricing problem, some economic models be used in spectrum leasing to meet PU revenue and SU communication quality. Since the propagation of radio in space is attenuated, the spectrum can be multiplexed in different regions, which means that SUs in different regions can share the spectrum. Since the spectrum usage of the PU is dynamic, offline auctions such as VCG cannot be used. An online auction mechanism based on the above characteristics is used to efficiently allocate spectrum resources (Sodagari et al. 2011). This online auction mechanism has multiple winners and the ability to complete auctions quickly. More information about spectrum

auctions can be found in the literature (Pastircak et al. 2014).

Key Applications

Dynamic spectrum access improves the efficiency of spectrum resources.

Future Directions

According to the current trend, it is a research direction for dynamic spectrum access to achieve smarter by deep learning. In addition, security issues, more accurate spectrum sensing, and applications towards 5G are also research hotspots for dynamic spectrum access.

Cross-References

► [TV White Space](#)

References

- Akyildiz Ian F et al (2006) NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey. *Comput Netw* 50(13):2127–2159
- Cao B, Zhang Q, Mark JW (2016) Introduction. In: Cooperative cognitive radio networking. Springer International Publishing, Switzerland, pp 4–10
- FCC (2010) Second memorandum opinion and order. http://hraunfoss.fcc.gov/edocs_public/attachmatch/FC_C-10-174A1.pdf
- Goldsmith A et al (2009) Breaking Spectrum gridlock with cognitive radios: an information theoretic perspective. *Proc IEEE* 97(5):894–914
- Khattab A, Perkins D, Bayoumi M (2013) Cognitive radio networks. Springer, New York
- Mitola J (2000) Cognitive radio: an integrated agent architecture for software defined radio. Royal Institute of Technology, PhD thesis
- Pastircak J, Gazda J, Kocur D (2014) A survey on the spectrum trading in dynamic spectrum access networks. In: Proceedings IEEE ELMAR 2014, Zadar, Croatia, pp 1–4
- Prasad RV et al (2008) Cognitive functionality in next generation wireless networks: standardization efforts. *IEEE Commun Mag* 46(4):72–78

- Ramani V, Sharma SK (2017) Cognitive radios: a survey on spectrum sensing, security and spectrum handoff. *China Commun* 14(11):185–208
- Sodagari S, Attar A, Bilen SG (2011) On a truthful mechanism for expiring Spectrum sharing in cognitive radio networks. *IEEE J Sel Areas Commun* 29(4): 856–865

Dynamic Spectrum Management

- [Dynamic Spectrum Access Through Cognitive Radio Networking](#)

E

Economic Efficiency

- [Efficiency and Pareto Optimality](#)

Edge Caching

- [Wireless Edge Caching](#)

Edge Computing

- [Architecture and Data Management for Smart Community Information Platform](#)
- [Data-Driven Edge Computing](#)
- [Edge Computing for Real-Time Video Stream Analytics](#)
- [Fog Computing for Intelligent Mobile and IoT Networks](#)
- [Mobile Edge Computing: Low Latency and High Reliability](#)

Edge Computing and Networking

- [Edge Data Centers in Fog Computing](#)

Edge Computing for Real-Time Video Stream Analytics

Miao Hu¹, Lei Zhuang¹, Di Wu¹, Zhichuang Huang¹, and Han Hu²

¹School of Data and Computer Science and Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China

²School of Information and Electronics, Beijing Institute of Technology, Beijing, China

Synonyms

[Edge computing](#); [Video analytics](#)

Definitions

Edge computing is a method of optimizing cloud computing systems by taking the control of computing applications, data, and services away from central nodes to the edge of the network. *Video analytics* is the capability of automatically analyzing video contents to detect and determine temporal or spatial events.

Historical Background

Satyanarayanan et al. (2009) first discussed the technical obstacles of mobile computing

and proposed a new architecture named as *cloudlet*, which is then unified into *edge cloud*. With an edge cloud, a mobile user exploits virtual machine (VM) or container technology to rapidly instantiate customized service and then uses that service over a wireless network. Generally, the edge cloud can be regarded as a trusted, resource-rich computer or cluster that's well-connected to the Internet and available for use by nearby mobile devices. Following the concept of edge computing, many works conducted case studies and presented challenges in the field of edge computing, especially for computation-intensive applications. For example, Ahmed and Ahmed (2016) discussed promising mobile edge computing scenarios, where video stream analytics is one killer application. Shi et al. (2016) highlighted some challenges and opportunities on edge computing in programmability, naming, data abstraction, service management, security and privacy, as well as optimization metrics that are worth future research and study.

A key benefit of utilizing edge computing is the capability to achieve high-throughput access, real-time processing guarantee, and energy saving. In the following, we will review some existing researches on video analytics system with edge computing from three aspects, processing efficiency, energy saving, and video characteristic extracting.

The concept of edge computing helps enhance communication bandwidth and reduce task processing time on real-time video stream analytics system design. Satyanarayanan et al. (2015) proposed a federated system of VM-based cloudlets that perform video analytics at the edge of the Internet, thus reducing the demand for ingress bandwidth into the cloud. To improve the efficiency of edge clouds, Salsano et al. (2017) introduced the architecture with dynamic deployment and composition of virtual functions and implemented a video streaming service exploiting mobile edge computing functionalities. To provide better scalability to the cloud-based solutions, Drolia et al. (2017) employed selective computation on the devices to reduce the need to off-load images to the cloud. In coordination

with edge servers, it uses prediction to prefetch parts of the trained classifiers used for recognition onto the devices and uses these smaller models to accelerate recognition on devices. Furthermore, Ren et al. (2018) proposed an edge computing-based object detection architecture to achieve distributed and efficient object detection via wireless communications for real-time surveillance applications. Overall, the benefits of real-time processing brought from edge computing lie in the high-throughput guarantee with edge clouds deployed in proximity.

In most of current mobile systems, many operations are conducted on the handheld devices, e.g., smartphones. As a result, developers worldwide are building increasingly complex applications that require ever-increasing amounts of computation resources and energy. Thus, the energy issue cannot be ignored, and fortunately this problem can be solved with the conception of edge computing. Cuervo et al. (2010) presented MAUI, a system that enables fine-grained energy-aware offload of mobile code to cloud. Kosta et al. (2012) proposed ThinkAir, a framework that makes it simple for developers to migrate their smartphone applications to cloud, which provides method-level computation offloading. Off-loading computation-intensive tasks to edge clouds can enhance energy efficiency, if the energy saving in computation processing is greater than the increased wireless transmission energy consumption.

To propose an efficient edge computing system for real-time video stream analytics, characteristics of video contents should also be taken into consideration. Bilal and Erbad (2017) highlighted some of the potentials and prospects of edge computing for interactive media and presented some preliminary works on how edge computing can be used to tackle various challenges. Ran et al. (2018) designed a framework that ties together front-end devices with more powerful back-end "helpers" to allow deep learning to be executed locally or remotely in the cloud/edge. They considered the complex interaction between model accuracy, video quality, battery constraints, network data usage, and network conditions to determine

an optimal offloading strategy. Wang et al. (2018) proposed an adaptive wireless video transcoding framework based on the emerging edge computing paradigm by deploying edge transcoding servers close to base stations. It is essential to efficiently extract video characteristics and then apply them into edge computing architecture design.

In summary, utilization of edge computing in video analytics can help save cost, bandwidth, energy, etc. Some special characteristics from video content or edge servers can be further exploited to enhance the system efficiency.

Foundations

In the environment of edge computing for video stream analytics, computation tasks can be off-loaded to edge servers to reduce latency and increase throughput. A task can be off-loaded to a nearby edge server, which returns the results back when the computation task is completed. To design a real-time video stream analytics in the edge computing architecture, we will face some new challenges:

- *Air interface management.* The wireless transmission will be severely affected by the interference among users. For example, the uplink data rate of a mobile user that chooses to off-load the computation to the edge server via a wireless channel can be modeled as

$$R = B \log \left(1 + \frac{P}{N_0 + I} \right) \quad (1)$$

where B is the channel bandwidth, P is the user transmission power, N_0 is the noise power, and I denotes the summarized interference power. It is important to schedule appropriate number of tasks to be served by edge server.

- *Frame extraction pattern.* As stated by many works, e.g., Lu et al. (2018), performing object detection in a video entails two main steps.

First, video frames must be extracted from the video and turned into images. To target objects with different dynamics within the video, the task issuer may request different frame extraction rates. For example, for an object moving at a high speed like a car, the rate should be high enough to not miss the object. For a slowly moving object like a pedestrian, a lower frame extraction rate can be used, which eases the computing burden. The setting of the frame extraction rate for object detection is important, which should be considered in the video analytics structure design.

- *Real-time scheduling design.* After the frames are extracted from the original video, object detection is performed on the images. The detection function may be performed locally on mobile devices, on edge servers, or on remote clouds or distributed among them. It is the task issuer's query requirements and the resource usages of mobile devices that should be considered in scheduling design.
- *Distributed edge colony.* The edge servers can be combined with each other to form a colony that can provide more powerful computing resources. In the realistic scenario, some mobile stations, e.g., the vehicles or even the mobile phones, can be regarded as the member of edge colony. The mobility of the edge server will introduce new problems, including "the dynamic edge computation capacity," "the information exchange between mobile nodes," etc.
- *Data consistency and integrity.* The captured video contents can be stored in the edge servers or remote cloud. Esposito et al. (2017) stated that it is important to cope with scalability, persistency, and reliability requirements and to adapt the infrastructure capacity to the exacting needs based on the amount of generated data. It is important to guarantee the data consistency and integrity in video analytics database design.

These challenges should be focused in applying edge computing technology into video analytics system design.

Applications

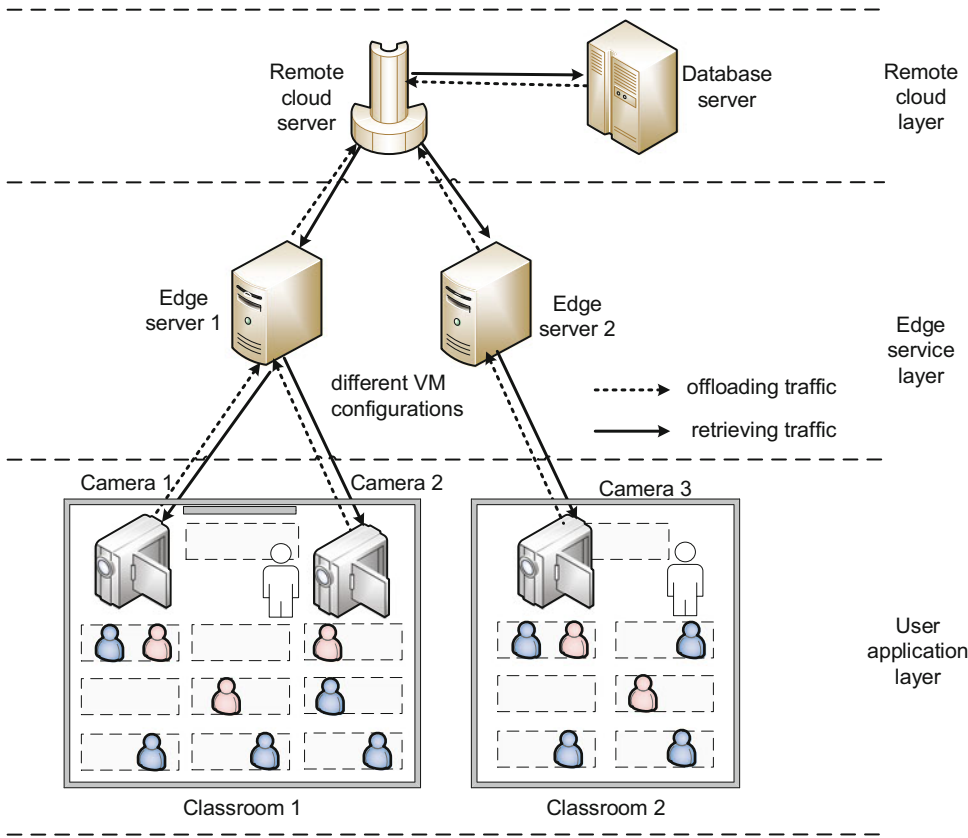
In education assistance applications, we design an edge computing framework for real-time video stream analytics, where the architecture can be found in Fig. 1. There are three key layers in our proposed edge computing architecture for real-time video stream analytics, including:

- **The user application layer.** In this layer, the camera will record the video shots for all the students' actions in the observed vision. After cumulating for a given period, the camera will off-load the captured video stream to the nearby edge server. The teacher with a handheld device can directly access the real-time processed student state information retrieved from the edge server.
- **The edge service layer.** In this layer, the edge server will conduct the video stream analytics

timely and send the required data back to the user (i.e., teacher). To enhance the tractability and efficiency of the edge computing architecture, the edge servers are combined into a colony. A leader edge node is elected to guarantee the load balance between different servers.

- **The cloud and database layer.** In this layer, an offline checking system is installed into the remote cloud server. After class, if the teacher has the interests to review the students' state of learning, he can check from the remote cloud. The video contents, in a specific period, can be stored in database for offline check.

The three layers should be jointly designed to construct a complete and efficient video analytics system with edge computing.



Edge Computing for Real-Time Video Stream Analytics, Fig. 1 An edge computing architecture for real-time video analytics in education assistance system

Future Directions

Edge computing paves a new way on video stream analytics, and the following up work possibly includes:

- *Multiview video stream analytics.* As shown in the classroom 1 of Fig. 1, the students' activities can be observed in different points of view. The video stream analytics results are overlapped with each other, which motivates a more efficient edge service schedule among them.
- *Multimodal information fusion.* Many multimodal information can be captured from the video stream, e.g., the face recognition information from the image and the voice information from the audio data. These information can be integrally learned to enhance the system performance. For instance, if the voice information from the acoustics is not detected, it might imply that the teacher is not teaching. In this case, the video analytics tasks might not be that urgent, which can be off-loaded to remote cloud.
- *Interactive video system.* In our previously discussed system, the video contents can only be uploaded to the edge servers, while user can receive this information afterward. In realistic word, a more interactive system is preferred, e.g., when the students who are observed can also receive analytics alerts, a more refined teaching ecosystem can be obtained.

Researches on these directions will further enhance system performance, which can be regarded as an efficient supplement to current video analytics architecture.

Cross-References

► Edge Computing

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grants 61802452, 61572538, Guangdong Special Support Program under Grant 2017TX04X148, the Fundamental Research Funds for the Central Universities under

Grant 17LGJC23, and the China Postdoctoral Science Foundation under Grant 2018M631025

References

- Ahmed A, Ahmed E (2016) A survey on mobile edge computing. In: Proceedings of 10th international conference on intelligent systems and control, Coimbatore, pp 1–8
- Bilal K, Erbad A (2017) Edge computing for interactive media and video streaming. In: Proceedings of 2nd international conference on fog and mobile edge computing, Valencia, pp 68–73
- Cuervo E et al (2010) MAUI: Making smartphones last longer with code offload. In: Proceedings of the 8th international conference on Mobile systems, applications, and services, San Francisco, pp 49–62
- Drolia U et al (2017) Precog: prefetching for image recognition applications at the edge. In: Proceedings of the 2nd ACM/IEEE symposium on edge computing, San Jose, pp 1–13
- Esposito C et al (2017) Challenges of connecting edge and cloud computing: a security and forensic perspective. *IEEE Cloud Comput* 4(2):13–17
- Kosta S et al (2012) ThinkAir: dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In: Proceedings of the 31st annual IEEE international conference on computer communications, Orlando, pp 945–953
- Lu Z et al (2018) A computing platform for video crowdprocessing using deep learning. In: Proceedings of the 37th annual IEEE international conference on computer communications, Honolulu, pp 1–9
- Ran X et al (2018) DeepDecision: a mobile deep learning framework for edge video analytics. In: Proceedings of the 37th annual IEEE international conference on computer communications, Honolulu, pp 1–9
- Ren J et al (2018) Distributed and efficient object detection in edge computing: challenges and solutions. *IEEE Netw* 1–7. <https://doi.org/10.1109/MNET.2018.1700415>
- Salsano S et al (2017) Toward superfluid deployment of virtual functions: exploiting mobile edge computing for video streaming. In: Proceedings of 29th international teletraffic congress, Genoa, pp 48–53
- Satyanarayanan M et al (2009) The case for VM-based cloudlets in mobile computing. *IEEE Pervasive Comput* 8(4):14–23
- Satyanarayanan M et al (2015) Edge analytics in the internet of things. *IEEE Pervasive Comput* 14(2): 24–31
- Shi W et al (2016) Edge computing: vision and challenges. *IEEE Internet Things J* 3(5):637–646
- Wang D et al (2018) Adaptive wireless video streaming based on edge computing: opportunities and approaches. *IEEE Trans Serv Comput* 1–12. <https://doi.org/10.1109/MNET.2018.1700415>

Edge Data Centers in Fog Computing

Lu Hou¹ and Kan Zheng²

¹Intelligent Computing and Communications Lab (IC² Lab), Beijing University of Posts and Telecommunications (BUPT), Beijing, China

²Intelligent Computing and Communications (IC2) Lab, Wireless Signal Processing and Networks (WSPN) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Synonyms

Edge computing and networking

Definitions

Edge data centers (EDCs) in fog computing bring some functionalities and capacities from centralized data centers to the network edge in order to meet the requirements of near-the-edge data service subscribers.

Historical Background

Centralized DCs (CDCs) are hard to handle the latency-sensitive data streams and the burst since they are usually deployed far away from subscribers. Fog computing brings the capacity from centralized data centers to the edges, providing low-latency, context-aware, high-availability, and distributed data services. EDCs are also responsible for data compression, de-duplication, and desensitization to ensure efficiency and safety. Some researches focus on the architecture of EDCs in fog computing environment. Zhang et al. in Zhang (2016) proposed a three-layer architecture consists of end users, operators, and fog nodes and provided distributed decisions to make full use

of computing resources by Stackelberg games. The work in Puthal (2018) studies the load balancing in EDCs in the aim of providing secure and sustainable services, balancing workloads between fog nodes, and backing up to make EDCs fault tolerance. Furthermore, Tan (2017) gave a load prediction approach to improve the rate of resource utilization in EDCs. Some other researches focus on the proper integration of CDCs and EDCs and to extend their advantages to extremity. Wang et al. in Wang (2018) focused on requirements of the Internet of things (IoT) and came up with a three-layer architecture where EDCs lie at the middle. The latencies of services of this architecture are evaluated, compared with CDCs-only scenarios. Sajjad (2016) proposed an approach for the deployment of geographically distributed data applications in order to reduce latencies and bandwidth utilities. There are also researchers studied EDCs in a software-defined networking (SDN)-based perspective, for example, in Martinello (2017), the authors proposed a programmable architecture of EDCs in term with key technologies applied in the architecture. The main idea is to decouple functions between edge and center, making the whole architecture flexible and intelligent. Dominicini et al. provided their researches over the network functions virtualization (NFV) in EDCs, bringing efficient orchestrations for service-centric architectures in Dominicini (2017). In addition, applications based on EDCs are considered in several other literatures. Sharma (2017) and Teranishi (2017) studied to apply EDCs in Internet of things networks or in dynamic data processing applications.

Foundations

The architecture of the data centers combined with fog computing can be divided into three layers. Apart from access networks layer and centralized data centers layer, edge data centers, deployed with some computing and storage devices, serve as the middle layer. The key idea is to treat the CDCs as a pool with massive computing and storage resources while process-

ing the real-time data streams over the edges. This kind of coordination can bring huge advantages for data services. The EDCs are formed up with several fog nodes in an orchestrated and distributed manner. One challenge related to EDCs is that both the capacity and the coverage are limited in fogs, for the reason that efficient resource management is critical to high QoS. Several ways can help manage resources in EDCs, for instance, giving the ability of adaption and adjustment of capacity of EDCs. The EDCs may also face the drastic fluctuations of data loads brought by the movements of users and the types of applications. With the help of location information in fog computing and proper learning algorithms, the EDCs can adjust promptly along with the movements of subscribers. Load balancing can also be deployed in fog computing to reassign data streams to different EDCs in order to equilibrate loads among them and stabilize the response time for subscribers. Another advantage which can be brought by load balancing is the high availability of EDCs services since load balancing can ensure the consistency of services by redirect loads of failed fog nodes to the surviving ones in the fog computing clusters. The nature of geographical distribution enables high availability for EDCs since they can backup each other over malfunctions. Both CDCs and EDCs have their own benefits. In the perspective of CDCs, EDCs play as an integral part which is supposed to provide data streams that are processed. In addition, EDCs should provide a set of APIs for CDCs to hide the heterogeneity. The harmonized cooperation between CDCs and EDCs offers end users with huge capabilities and low latencies of data services. One efficient way of integration is introducing the paradigms of SDN and NFV. With the decoupling of control plane and data plane, EDCs can provide flexibility for the deployment of network functions in fog computing. As a consequence, intelligence can be brought to improve the expandability and robustness. On the other hand, SDN architecture distinct nothing between CDCs and EDCs, they can function in a conjunct way. Thus, data services can be deployed and managed in an easy way.

Key Applications

Edge data centers are used in every environment where low-latency services are required to be moved from data centers to the fog, for example, content delivery, Internet of things, and intelligent transportation systems.

Future Directions

Unlike CDCs, EDCs are under the intimidation of viciousness. Security is vital in the fog environments. Therefore, EDCs needs to be deployed with critical arrangement of security. On the other hand, the drastic changes of physical channels bring another challenge to the application of data centers in fog computing.

Cross-References

- [Data Center Networking \(DCN\): Infrastructure and Operation](#)
- [Data Center Network Virtualization](#)
- [Resource Allocation](#)

References

- Dominicini C, Vassoler G, Meneses L, Villaca R, Ribeiro M, Martinello M (2017) VirtPhy: fully programmable NFV orchestration architecture for edge data centers. *IEEE Trans Netw Serv Manag* 14(4):817–830
- Jarray A, Salazar J, Karmouch A, Elias J, Mehaoua A (2014) Column generation based-approach for IaaS aware networked edge data-centers. In: *IEEE Globecom workshops (GC Wkshps)*
- Martinello M, Liberato A, Beldachi A, Kondepu K, Gomes R, Villaca R, Ribeiro M, Yan Y, Hugues-Salas E, Simeonidou D (2017) Programmable residues defined networks for edge data centres. In: *IEEE International conference on network and service management (CNSM)*
- Puthal D, Obaidat M, Nanda P, Prasad M, Mohanty S, Zomaya A (2018) Secure and sustainable load balancing of edge data centers in fog computing. *IEEE Commun Mag* 56(5):60–65
- Sajjad H, Danniswara K, Al-Shishtawy A, Vlassov V (2016) SpanEdge: towards unifying stream processing over central and near-the-edge data centers. In: *IEEE/ACM symposium on edge computing (SEC)*

- Sharma S, Wang X (2017) Live data analytics with collaborative edge and cloud processing in wireless IoT networks, vol 5. IEEE Access
- Teranishi Y, Kimata T, Yamanaka H, Kawai E, Harai H (2017) Dynamic data flow processing in edge computing environments. In: IEEE 41st annual computer software and applications conference (COMPSAC)
- Wang P, Chen X, Sun Z (2018) Performance modeling and suitability assessment of data center based on fog computing in smart systems, vol 6. IEEE Access
- Tan C, Klein C, Elmroth E (2017) Location-aware load prediction in edge data centers. In: IEEE Second international conference on fog and mobile edge computing (FMEC)
- Zhang H, Xiao Y, Bu S, Niyato D, Yu R, Han Z (2016) Fog computing in multi-tier data center networks: a hierarchical game approach. IEEE International Conference on Communications (ICC)

Edge Forensics

► Internet of Things (IoT) Forensic Science

Effective Utilization of Unlicensed Spectrum

Yu Gu and Qimei Cui
Beijing University of Posts and
Telecommunications, Beijing, China

Synonyms

Licensed-assisted access (LAA); LTE-unlicensed (LTE-U); Quality-of-service (QoS)

Definition

Licensed-assisted access (LAA) or LTE-unlicensed (LTE-U), is a promising technology to offload dramatically increasing cellular traffic to unlicensed bands, exploiting Listen-before-talk (LBT). The main goal is to achieve effective utilization of unlicensed spectrum and fair coexistence with legacy systems such as IEEE 802.11 WiFi.

Historical Background

In the recent years, the increasing prevalence of smart hand-set device (e.g., mobiles, laptops) has triggered an explosive growth of mobile data stemming. Forecasts from Cisco show the traffic demand is expected 1000 times as much traffic by 2020 (Cisco Visual Networking Index 2015). The scarcity of spectrum is becoming the bottleneck to further boost the capacity of wireless communications (Zhang et al. 2015a). To fundamentally break through this predicament, academia and industry paid attention to more available spectrum opportunities, such as industry, science, and medicine (ISM) bands in 2.4 GHz, the unlicensed national information infrastructure (U-NII) bands in 5.8 GHz, etc. An emerging technology to supplementary utilization of the unlicensed spectrum, called licensed-assisted access (LAA) or LTE-unlicensed (LTE-U), has been launched into the standardization by Third Generation Partnership Project (3GPP) (3GPP 2015). The protocol also remains the strongest candidates for the upcoming 5G cellular communication standard on the exploitation of the unlicensed spectrum (RP-170828). Originally designed to off-load excessive traffic for the downlinks from the licensed bands, the protocols have also been considered for the uplink applications in the unlicensed band (3GPP 2015). Different from Wi-Fi APs, the underlying network infrastructure can coordinate between LAA stations in the unlicensed bands and base stations in licensed bands. Signaling, such as channel state information (CSI) and channel quality indicator (CQI), can also be instantly fed back from user terminals to LAA stations, through the licensed band, thereby facilitating fast power control and user scheduling of LAA in the unlicensed bands.

The LAA technology is still in its infancy, and there are a few critical challenges arising in network coexistence among different radio access technologies, spectrum sharing, and access (Yu et al. 2016). The major challenge is how to guarantee the fair and effective coexistence between LAA-enabled small base stations

(SBSs) and Wi-Fi. Then, future 5G systems will support a variety of applications with wide-ranging QoS requirements, including VoIP, virtual reality, multimedia streaming, machine-type communications, as well as traditional web browsing and file transfer. The random arrived traffic and the random access mechanism of LAA become an obstacle to guarantee QoS. Furthermore, the new LAA procedures may have impacts on energy consumption of mobile devices due to the extra energy used for channel detection and packet collision. Lastly, due to the time-variant wireless channel conditions and the dynamic variation of the number of Wi-Fi nodes (e.g., Wi-Fi stations, Wi-Fi APs) in the overlapping areas, LAA-enabled SBSs need a dynamic spectrum access mechanism to switch to an unlicensed channel with very-low-level interference and leverage the traffic between the licensed and unlicensed bands.

As for the coexistence of SBSs and Wi-Fi, two kinds of specifications are proposed: frame-based mechanism (FBM) where SBS is activated at periodic cycles on unlicensed band and load-based mechanism (LBM) where SBS competes for the unlicensed channel using listen-before-talk (LBT) and backoff procedure like Wi-Fi (3GPP 2015; [Broadband Radio Access Networks \(BRAN\)](#)). Zhang et al. (2015b), Al-Dulaimi et al. (2015), and Ratasuk et al. (2014) design coexistence mechanisms, such as an almost blank sub-frame (ABS) scheme, an interference avoidance scheme (Zhang et al. 2015b), and adaptive listen-before-talk (LBT) mechanism (Zhang et al. 2015b; Ratasuk et al. 2014). To improve the system throughput, Rupasinghe and Güvenç (2015) propose a Q-Learning based dynamic duty cycle selection technique for configuring LTE transmission gaps. Almeida et al. (2013) propose a coexistence scheme by allocating the almost blank sub-frames to improve the Wi-Fi throughput per user up to 50 times. In Han et al. (2016), an LBT-based MAC protocol is developed, where the transmission durations of LAA devices are optimized to maximize the system throughput. In Voicu et al. (2016), a comparison study is conducted between the approaches using duty cycle and LBT and revealed the effectiveness of

LBT in the case of strong interference. The above works mainly focus on the design of coexistence mechanisms or improving the system throughput, without considering the QoS issue and the dynamic varying of wireless environment.

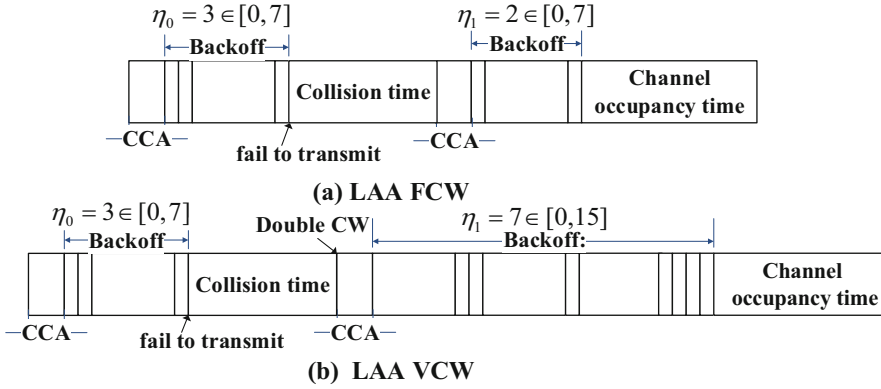
A few number of works have studied on QoS or energy efficiency (EE) requirements of SBS in the unlicensed band. Yin et al. (2016) develop a power allocation algorithm to obtain pareto-optimal between minimization of interference in the licensed band and collision in the unlicensed band by the weighted Tchebycheff method while satisfying the rate requirements of users. Chen et al. (2016) first investigate joint licensed and unlicensed resource allocations to maximize the EE through Nash bargaining when LAA systems adopt a FBM method.

Foundations

As specified in 3GPP TR 36.889 (3GPP 2015), two different channel access schemes are considered for LAA, i.e., LBT with random backoff in a fixed CW and LBT with random backoff in an exponentially increasing CW, as illustrated in Fig. 1a, b, respectively. These two methods are referred to as FCW (Fixed CW) and VCW (Variable CW), respectively.

In the case of FCW, a LAA-BS senses the unlicensed band for a predefined period, termed “channel clear assess (CCA),” whenever it has packets to transmit. If the channel is free over CCA, the LAA-BS sets an integer backoff timer randomly and uniformly within a fixed CW $[0, W_L]$, where W_L is the initial CW size. The backoff timer counts down one per timeslot. It freezes if the channel is busy and does not resume until the channel is sensed free for CCA again. Once the backoff timer turns to zero, a (re)transmission is triggered. If the (re)transmission is collided (i.e., no acknowledgment (ACK) is returned), the LAA-BS resets the backoff timer within $[0, W_L]$ to retransmit the packet.

In the case of VCW, exponential backoff is adopted on top of FCW, where the CW doubles, each time the (re)transmission of the LAA-BS



Effective Utilization of Unlicensed Spectrum, Fig. 1 Medium access mechanism for unlicensed spectrum: (a) LAA FCW and (b) LAA VCW, where the initial CW

of LAA is 8, and η_j is the number of backoff timeslots in response to the j -th collided (re)transmission, $j = 0, \dots, K_L - 1$

is collided (i.e., no ACK is returned). K_L is the maximum number of retransmissions per packet, after which the CW is reset to W_L . In this sense, VCW is analogous to the distributed coordination function (DCF) of Wi-Fi (Bianchi 2000). However, the LAA-BS resets the CW to W_L , after a collision-free (re)transmission (i.e., neither ACK nor non-ACK is returned), as opposed to the DCF. This is due to the fact that the LAA-BS can exploit OFDMA to multiplex signals for multiple LAA-UEs. It is possible that only some of the LAA-UEs succeed and return ACKs after a collision-free (re)transmission (3GPP 2015). Resetting the CW can prevent the CW from continuously enlarging and staying large.

Unfortunately, it is difficult to preserve reliability in the unlicensed spectrum which is primarily dominated by uncoordinated transmissions of contention-based wireless devices. Erratic collisions between the uncoordinated transmissions can breach the delay requirements of traffic and even cause significant packet losses. The ultra-dense deployment of networks could further deteriorate the reliability in the unlicensed spectrum. Moreover, there is typically no policy to regulate the deployment of wireless transmitters in the unlicensed band, given the distributed nature of the networks and devices. The literatures Cui et al. (2017, 2018) have propose a novel effective capacity to quantify the maximum consistent data rate of a link without violating the QoS in

unlicensed spectrum. Potential practical applications of the new measure are also discussed.

QoS-aware Power Control

The maximization of the effective capacity in unlicensed spectrum can be formulated by optimizing the power of a transmitter concurrently sending data to multiple receivers using orthogonal frequency-division multiple access (OFDMA) techniques. By exploiting the concavity of the effective capacity, the problem can be reformulated to be a convex optimization problem which can be straightforwardly solved by using standard convex techniques, such as interior-point method and sub-gradient method.

QoS-aware Bandwidth Allocation

The effective capacity can also be used to facilitate the bandwidth allocation in the case that a LAA transmitter serves multiple receivers with nontrivial QoS requirements. Exploiting the concavity of the effective capacity with respect to the instantaneous transmit rate, the bandwidth allocation can be formulated as a convex optimization problem which is readily solved by using standard convex techniques.

Other Applications

Another application example of the effective capacity is to maximize the effective energy efficiency of transmitters in unlicensed spectrum.

This is an extension of the above effective capacity maximization, since the effective energy efficiency of a transmitter is defined to be the ratio of its effective capacity to its transmit power. By exploiting the aforementioned concavity of the effective capacity again, as well as fractional programming techniques, the maximization of the effective energy efficiency can be formulated to be a parametric convex problem and recursively solved by using the KKT conditions and the Dinkelbach's method.

Cross-References

- [Licensed-Assisted Access \(LAA\)](#)
- [LTE-Unlicensed \(LTE-U\)](#)
- [Quality-of-Service \(QoS\)](#)
- [Resource Allocation](#)

References

- 3GPP (2015) Feasibility study on licensed-assisted access to unlicensed spectrum, 3GPP TR 36.889 V13.0.0, June 2015
- Al-Dulaimi A, Al-Rubaye S, Ni Q, Sousa E (2015) 5G communications race: pursuit of more capacity triggers LTE in unlicensed band. *IEEE Veh Technol Mag* 10(1):43–51
- Almeida E, Cavalcante AM, Paiva RCD, Chaves FS, Abinader FM, Vieira RD, Choudhury S, Tuomaala E, Doppler K (2013) Enabling LTE/WiFi coexistence by LTE blank subframe allocation. In: *Proceedings of IEEE International Conference on Communications (ICC)*, June 2013, pp 5083–5088
- Bianchi G (2000) Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J Sel Areas Commun* 18(3):535–547
- Broadband Radio Access Networks (BRAN). 5 GHz high performance RLAN, ETSI EN 301 893
- Chen Q, Yu G, Yin R, Maaref A, Li GY, Huang A (2016) Energy efficiency optimization in licensed-assisted access. *IEEE J Sel Areas Commun* 34(4):723–734
- Cisco Visual Networking Index (2015) Global mobile data traffic forecast update, 2014–2019, Feb 2015
- Cui Q, Gu Y, Ni W, Liu RP (2017) Effective capacity of licensed-assisted access in unlicensed spectrum for 5G: from theory to application. *IEEE J Sel Areas Commun* 35(8):1754–1767
- Cui Q, Gu Y, Ni W, Zhang X, Tao X, Zhang P, Liu RP (2018) Preserving reliability to heterogeneous ultra-dense distributed networks in unlicensed spectrum. *IEEE Commun Mag* 56(6):72–78
- Han S, Liang YC, Chen Q, Soong BH (2016) Licensed-assisted access for LTE in unlicensed spectrum: a MAC protocol design. *IEEE J Sel Areas Commun* 34(10):2550–2561
- Ratasuk R, Mangalvedhe N, Ghosh A (2014) LTE in unlicensed spectrum using licensed-assisted access. In: *Proceedings of IEEE Globecom Workshops (GC Wkshps)*, Dec 2014, pp 746–751
- RP-170828. New SID on NR-based access to unlicensed spectrum, 3GPP TSG RAN Meeting #75, Dubrovnik, 6–9 Mar
- Rupasinghe N, Güvenç İ (2015) Reinforcement learning for licensed-assisted access of LTE in the unlicensed spectrum. In: *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, Mar 2015, pp 1279–1284
- Voicu AM, Simi L, Petrova M (2016) Inter-technology coexistence in a spectrum commons: a case study of Wi-Fi and LTE in the 5-GHz unlicensed band. *IEEE J Sel Areas Commun* 34(11):3062–3077
- Yin R, Yu G, Maaref A, Li GY (2016) A framework for co-channel interference and collision probability trade-off in LTE licensed-assisted access networks. *IEEE Trans Wirel Commun* 15(9):6078–6090
- Yu G, Li GY, Wang LC, Maaref A, Lee J, Lopez-Perez D (2016) Guest editorial: LTE in unlicensed spectrum. *IEEE Wirel Commun* 23(6):6–7
- Zhang R, Wang M, Cai LX, Zheng Z, Shen X, Xie LL (2015a) LTE-unlicensed: the future of spectrum aggregation for cellular networks. *IEEE Wirel Commun* 22(3):150–159
- Zhang H, Chu X, Guo W, Wang S (2015b) Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum. *IEEE Commun Mag* 53(3):158–164

Efficiency and Pareto Optimality

Xiaowen Gong¹, Lei Yang², Xu Chen³, and Junshan Zhang⁴

¹Auburn University, Auburn, AL, USA

²University of Nevada Reno, Reno, NV, USA

³School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

⁴Arizona State University, Tempe, AZ, USA

Synonyms

[Economic efficiency](#); [Pareto efficiency](#); [Social optimality](#)

Definitions

An economic system is efficient if the resource production and allocation maximizes the total payoff of all agents in the system. More generally, a system is efficient if the system state maximizes a desired performance metric (such as the total utility of all entities in the system) given the inputs to the system.

An economic system is Pareto-optimal if there does not exist alternative resource production and allocation that improves at least one agent's payoff without reducing any agent's payoff. More generally, a system is Pareto-optimal if there does not exist an alternative system state that makes at least one preference metric (such as some entity's utility) better off without making any preference metric worse off.

Historical Background

The concept of Pareto efficiency is named after Vilfredo Pareto who was an Italian engineer and economist. He used the concept in his studies of economic efficiency and income distribution. Pareto efficiency is closely related to two fundamental theorems. The first fundamental theorem was first demonstrated graphically by economist Abba Lerner and first mathematically by economists Kenneth Arrow and Gérard Debreu. It states that given certain assumptions, competitive markets produce Pareto efficient outcomes (Hindriks and Myles 2013). This captures Adam Smith's "invisible hand" hypothesis, namely, that competitive markets tend toward an efficient allocation of resources (Mas-Colell et al. 1995). The second fundamental theorem states that given further restrictions, any Pareto efficient outcome can be supported as a competitive market equilibrium (Hindriks and Myles 2013).

Key Points

The efficiency of a system is usually measured in terms of the social welfare, which is the total utility of all entities in the system. It is often

a desirable objective for system designers and has been used in many applications including economics and engineering. The metric of efficiency is in stark contrast to fairness, which is often concerned with maximizing the minimum utility among all entities in the system. As a result of this difference, an efficient system can be quite different from a fair system and may not be fair.

Pareto optimality is a weaker notion than efficiency, in the sense that an efficient system (measured by social welfare) must be Pareto-optimal, but the converse may not hold. Pareto optimality is considered a minimal notion of "efficiency" that does not necessarily result in a socially desirable state or fairness among the entities. As a result, it is almost always one of the desired objectives for system design. The concept has been applied in various fields such as economics and engineering. Simply put, a system is not Pareto-optimal if there is an alternative system state, in which further improvements can be made to at least one agent's payoff without sacrificing any other agent's payoff. If there is a state transfer that satisfies this condition, it is called a "Pareto improvement." The set of all Pareto-optimal system states constitute the Pareto frontier. A system is weakly Pareto-optimal if there does not exist an alternative system state that makes every entity better off.

The bargaining solution is an important concept that satisfies Pareto optimality. It is the solution to the bargaining problem, which is concerned with how agents share a benefit that they can jointly generate by cooperation. A prominent example of the bargaining solution is the Nash bargaining solution (Osborne and Rubinstein 1994). It is the unique solution to a two-person bargaining problem that has desired properties by satisfying the axioms of scale invariance, symmetry, independence of irrelevant alternatives, and Pareto optimality.

Applications in Wireless Networks

In the following, three example applications are presented to illustrate how efficiency and Pareto optimality are used to design, model, and analyze

wireless networks. The first two applications focus on finding the Pareto-optimal solution and its relation with the system-wide efficient (i.e., socially optimal) solution; the last application is concerned with achieving the system-wide efficient solution.

Social Group Utility Maximization-Based Pareto-Optimal Pseudonym Change for Mobile Location Privacy

To protect location privacy in a location-based services (LBS), mobile users in physical proximity can simultaneously change their pseudonyms used in the LBS (Beresford and Stajano 2003), so that their pseudonyms cannot be distinguished from each other after the pseudonym change. Since pseudonym change usually incurs considerable overhead (e.g., service interruption (Beresford and Stajano 2003; Freudiger et al. 2009)), users need adequate incentive (e.g., privacy gain) to participate. On the other hand, with the rapid growth of online social networks (e.g., Facebook, Twitter), people's behavior tend to be altruistic to their social "friends" (friends, family, colleagues, etc.), as they care about their social friends' welfare. Therefore, mobile users' socially altruistic behavior can be exploited for pseudonym change to improve their location privacy.

Based on this motivation, mobile users' socially altruistic behavior is exploited in Gong et al. (2017) to encourage them to participate in pseudonym change based on a social group utility maximization (SGUM) framework developed in Chen et al. (2016). Under the SGUM framework, each user $i \in \mathcal{N}$ makes a decision a_i on whether to participate in pseudonym change ($a_i = 1$) or not ($a_i = 0$). The privacy gain of a user participating in pseudonym change depends on how many users in its anonymity set $\mathcal{N}_{i-}$ also participate, where $\mathcal{N}_{i-}$ depends on users' physical locations. Each user i 's individual utility u_i is given by $u_i(a_i, \mathbf{a}_{-i}) \triangleq a_i \left(\sum_{j \in \mathcal{N}_{i-}} a_j - c_i \right)$ where $\mathbf{a}_{-i}$ is all users' actions except for user i . If a user participates, its individual utility is its privacy gain $\sum_{j \in \mathcal{N}_{i-}} a_j$ minus its participation cost c_i ; otherwise, it is zero. To take into account users' socially

altruistic behavior, each user i aims to maximize its *social group utility* given by $f_i(a_i, \mathbf{a}_{-i}) \triangleq u_i(a_i, \mathbf{a}_{-i}) + \sum_{j \in \mathcal{N}_{i+}} s_{ij} u_j(a_j, \mathbf{a}_{-j})$ where $s_{ij} \in (0, 1]$ denotes user i 's *social tie* with user j , which quantifies how much user i cares about user j .

Similar in spirit to the Nash equilibrium (Osborne and Rubinstein 1994) of a standard noncooperative game, the following concept applies to the SGUM game (Chen et al. 2016). A strategy profile $\mathbf{a}^{SNE} = (a_1^{SNE}, \dots, a_n^{SNE})$ is a *socially aware Nash equilibrium* (SNE) for the socially altruistic pseudonym change game (SA-PCG) if no user can improve its social group utility by unilaterally changing its strategy, i.e., $a_i^{SNE} = \arg \max_{a_i \in \{0,1\}} f_i(a_i, \mathbf{a}_{-i}), \forall i \in \mathcal{N}$.

The *social welfare* of the system is the total individual utility of all users given by $v(\mathbf{a}) \triangleq \sum_{i \in \mathcal{N}} u_i(\mathbf{a})$. A strategy profile $\mathbf{a}^{SO} = (a_1^{SO}, \dots, a_n^{SO})$ is *social optimal* if it achieves the maximum social welfare among all strategy profiles, i.e., $v(\mathbf{a}^{SO}) \geq v(\mathbf{a}), \forall \mathbf{a}$.

A strategy profile $\mathbf{a}^{PO} = (a_1^{PO}, \dots, a_n^{PO})$ is *Pareto-optimal* if there does not exist a *Pareto-superior* profile $\mathbf{a}' = (a'_1, \dots, a'_n)$ such that no user achieves a lower individual utility, while at least one user achieves a higher individual utility, i.e., $u_i(a'_i, \mathbf{a}'_{-i}) \geq u_i(a_i^{PO}, \mathbf{a}_{-i}^{PO}), \forall i \in \mathcal{N}$ with at least one strict inequality.

As a benchmark, in contrast to SA-PCG, a socially oblivious pseudonym change game (SO-PCG) is one in which all social ties $s_{ij}, \forall i \neq j$ are 0.

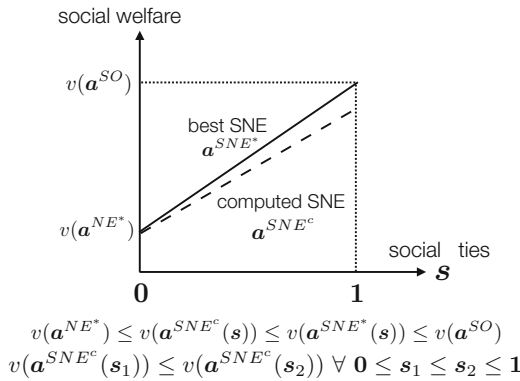
To quantify the system efficiency of the SNE, consider the social welfare of the best SNE $\mathbf{a}^{SNE*}$, which is the SNE of the highest social welfare among all the SNEs. Let $\mathcal{N}_{i+}$ denote the set of users whose anonymity set includes user i . It is shown in Gong et al. (2017) that the performance gap between the maximum social welfare among all SNEs and the optimal social welfare is upper bounded by $\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}_{i+}} (1 - s_{ij})$. This shows that the performance gap decreases as social ties increase. In particular, when users are socially oblivious (i.e., $s_{ij} = 0, \forall i \neq j$), the performance gap reaches the maximum, and the best SNE for the SA-PCG degenerates to the best SNE for the

SO-PCG; when users are fully altruistic (i.e., $s_{ij} = 1, \forall i \neq j$), the performance gap becomes 0, and the best SNE degenerates to the social optimal strategy profile.

An efficient algorithm has been developed in Gong et al. (2017) to find a Pareto-optimal SNE $\mathbf{a}^{SNE^c}$. It can be shown that the social welfare of the SNE $\mathbf{a}^{SNE^c}$ is no less than that of the best SNE for the SO-PCG.

It is shown in Gong et al. (2017) that, for the SA-PCG, when social ties increase (i.e., $s'_{ij} \geq s_{ij}, \forall i \neq j$), the set of participating users at the SNE $\mathbf{a}^{SNE^c}$ grows (i.e., $\mathcal{N}_1(\mathbf{a}^{SNE'}) \supseteq \mathcal{N}_1(\mathbf{a}^{SNE^c})$) and the social welfare of the SNE $\mathbf{a}^{SNE^c}$ increases (i.e., $v(\mathbf{a}^{SNE'}) \geq v(\mathbf{a}^{SNE^c})$), as illustrated in Fig. 1). This shows that $\mathbf{a}^{SNE^c}$ is monotonically “expanding” with respect to social ties. Intuitively, a user with larger social ties with other users is more likely to participate in favor of its social group utility, even at the cost of obtaining a negative individual utility. Therefore, when social ties become larger, more users participate at the SNE. Furthermore, as each additional participating user increases the social welfare, the social welfare of the SNE also increases.

It can be shown that the social welfare of the SNE $\mathbf{a}^{SNE^c}$ for the SA-PCG is no less than that of the best SNE for the SO-PCG. This guarantees that the social welfare of the Pareto-optimal SNE is no less than the benchmark SNE for the SO-PCG.



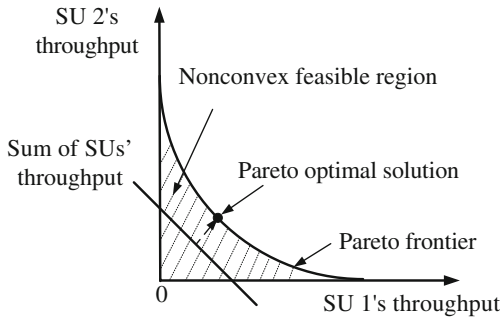
Efficiency and Pareto Optimality, Fig. 1 The computed SNE migrates monotonically from Nash equilibrium to social optimality

Pareto-Optimal Spectrum Access for Secondary Users in Dynamic Spectrum Sharing

Cognitive radio is expected to capture temporal and spatial “spectrum holes” in the spectrum white space and to enable spectrum sharing for secondary users (SUs). Pricing-based mechanisms have been explored as a promising approach for spectrum access, where PUs can dynamically trade unused spectrum to SUs. However, SUs may compete via random access for spectrum opportunities. In this case, it is interesting to study pricing-based spectrum access control in cognitive radio networks, where PUs sell the temporarily unused spectrum and SUs compete via random access for such spectrum opportunities.

Consider a monopoly PU market model, where the PU’s unused spectrum is fixed. Pricing-based dynamic spectrum access based on slotted Aloha has been studied in Yang et al. (2013), aiming to characterize the optimal pricing strategy to maximize the PU’s revenue. The prices for each SU to access spectrum consist of usage price p_i and flat price g_i . Given the prices, each SU will optimize its demand by maximizing its net utility. Let $U_i(s_i)$ denote the utility function and s_i be the throughput of SU i . Under random access, s_i represents the expected number of successfully transmitted packets of SU i in a period of c and can be written as $s_i = cz_i \prod_{j \neq i} (1 - z_j)$, where z_i is the spectrum access probability of SU i . Thus, the corresponding demand function of SU i can be derived as $d_i(p_i) = U_i'^{-1}(p_i)$ if $g_i \leq U_i(s_i) - p_i s_i$ and $d_i(p_i) = 0$ (i.e., SU i would not transmit when the net utility is negative), otherwise. The PU’s revenue maximization (RM) problem is to maximize $\sum_{i \in \mathcal{M}} (g_i + p_i s_i)$ subject to the constraints $s_i \leq cz_i \prod_{k \neq i} (1 - z_k), \forall i \in \mathcal{M}$, $s_i = d_i(p_i), \forall i \in \mathcal{M}$, and $U_i(s_i) - g_i - p_i s_i \geq 0, \forall i \in \mathcal{M}$.

The RM problem is nonconvex and therefore it is in general difficult to find the global optimum. Observing that the global optimum of the RM problem is Pareto-optimal, it is reasonable to consider the Pareto frontier, i.e., the set consisting of all the Pareto-optimal solutions. The



Efficiency and Pareto Optimality, Fig. 2 Sketch of the Pareto-optimal solution: the case with two SUs

solution to the RM problem has the following properties:

1. It is in the Pareto frontier.
2. It is Pareto-optimal if and only if $\sum_{i \in \mathcal{M}} z_i = 1$.

For any Pareto-optimal allocation $d_i(p_i)$, $\forall i \in \mathcal{M}$, it holds that $\sum_{i \in \mathcal{M}} z_i = 1$. Therefore, it suffices to search for the solutions in the Pareto frontier, which can maximize the RM problem. Thus motivated, it is desirable to obtain a Pareto-optimal solution to the RM problem that maximizes the sum of SUs' throughput given the SUs' budget constraints, by confining the search space to the hyperplane $\sum_{i \in \mathcal{M}} d_i(p_i) = \kappa c$, where $\kappa \in [0, 1]$ denotes the *spectrum utilization percentage* under the allocation $d_i(p_i)$, $\forall i \in \mathcal{M}$. An example with two SUs is illustrated in Fig. 2.

When SUs' utility functions are α -fair utility functions, a Pareto-optimal solution to the RM problem has been found in Yang et al. (2013). Interestingly, the Pareto-optimal solution converges to the global optimal, as α goes to zero as shown in Yang et al. (2013). When $\alpha = 0$, the PU selects the SU with the highest budget and only allows that SU to access the channel with probability 1 to maximize its revenue.

Asymptotically Socially Optimal Distributed Channel Selection in Database-Assisted Spectrum Access

The most recent FCC ruling requires that TV white-space devices must rely on a geo-location

database to determine the spectrum availability (FCC September 23, 2010). In this architecture, the database is able to tell a white-space device (secondary users (SUs) of TV spectrum) vacant TV channels at a particular location. Motivated by the successful deployments of Wi-Fi over the unlicensed ISM bands, Chen and Huang (2013) consider an infrastructure-based white-space network, where there are multiple secondary access points (APs) operating on white spaces. A key challenge for such an infrastructure-based white-space network design would be that each AP must choose a proper vacant channel to operate in order to avoid severe interference with other APs.

Consider that all the APs try to maximize the system-wide throughput (i.e., efficiency) cooperatively. Such a cooperation is feasible when all the APs are owned by the same network operator. For example, the APs that are deployed in a university campus can coordinate to maximize the entire campus network throughput. Formally, the APs need to collectively determine the optimal channel selection profile a such that the system-wide throughput is maximized (problem P1), i.e., $\max_{a \in \Theta} \sum_{n=1}^N U_n(a)$, where $U_n(a)$ is the expected throughput of AP n under the channel selection profile a . Problem P1 is a combinatorial optimization problem of finding the optimal channel selection profile over the discrete solution space Θ . In general, such a problem is very challenging to solve exactly especially when the size of network is large (i.e., the solution space Θ is large).

A cooperative channel selection algorithm has been proposed in Chen and Huang (2013) that can approach the optimal system-wide throughput and thus achieve efficiency approximatively. To proceed, first write problem P1 into the following equivalent problem (problem P2): $\max_{(q_a: a \in \Theta)} \sum_{a \in \Theta} q_a \sum_{n=1}^N U_n(a)$, where q_a is the probability that channel selection profile a is adopted. Obviously, the optimal solution to problem P2 is to choose the optimal channel selection profile with probability one. It is known from Chen et al. (2010) that problem P2 can be approximated by the following convex optimization problem (problem P3): $\max_{(q_a: a \in \Theta)} \sum_{a \in \Theta} q_a \sum_{n=1}^N U_n(a) - \frac{1}{\gamma} \sum_{a \in \Theta} q_a$

$\log q_a$, where γ is the parameter that controls the approximation ratio. It can be seen that when $\gamma \rightarrow \infty$, problem P3 becomes exactly the same as problem P2. That is, when $\gamma \rightarrow \infty$, the optimal point a^* that maximizes the system throughput $\sum_{n=1}^N U_n(a)$ will be selected with probability one. A nice property of such an approximation in problem P3 is that the close-form solution can be obtained, which enables the distributed algorithm design later. More specifically, by the KKT condition (Boyd and Vandenberghe 2004), the optimal solution to problem P3 can be derived as

$$q_a^* = \frac{\exp\left(\gamma \sum_{n=1}^N U_n(a)\right)}{\sum_{a' \in \Theta} \exp\left(\gamma \sum_{n=1}^N U_n(a')\right)}.$$

Similarly to the spatial adaptive play in Young (2001) and Gibbs sampling in Kauffmann et al. (2007), a cooperative AP channel selection algorithm is designed in Chen and Huang (2013) by carefully coordinating APs' asynchronous channel selection updates to form a discrete-time Markov chain (with the system state as the channel selection profile a of all APs). Since each AP will be selected to update with a probability of $\frac{1}{N}$ and the selected AP will randomly choose a channel with a probability proportional to $\exp\left(\gamma \sum_{n=1}^N U_n(a)\right)$, then if $a' \in \Lambda_a$, the probability that the Markov chain transits from state a to a' is given as $q_{a,a'} = \frac{1}{N} \frac{\exp\left(\gamma \sum_{n=1}^N U_n(a'_n, a_{-n})\right)}{\sum_{a' \in \mathcal{M}_n} \exp\left(\gamma \sum_{n=1}^N U_n(a', a_{-n})\right)}$.

Otherwise, it holds that $q_{a,a'} = 0$. It is shown in Chen and Huang (2013) that the cooperative AP channel selection algorithm induces a time-reversible Markov chain with the unique stationary distribution as given above.

The social optimal point that maximizes the system-wide throughput can be approached by setting $\gamma \rightarrow \infty$ in the cooperative AP channel selection algorithm. However, in practice only a finite value of γ can be implemented such that $\exp\left(\gamma \sum_{n=1}^N U_n(a)\right)$ does not exceed the range of the largest predefined real number on a computer. Let $\bar{S} = \sum_{a \in \Theta} q_a^* \sum_{n=1}^N U_n(a)$ be the expected potential by the proposed algorithm and $S^* = \max_{a \in \Theta} \sum_{n=1}^N U_n(a)$ be the global optimal

potential. It is shown in Chen and Huang (2013) that, when a large enough γ is adopted, the performance gap between $\bar{S}$ and S^* is very small, thus achieving efficiency approximately. For the cooperative AP channel selection algorithm, it holds that $0 \leq S^* - \bar{S} \leq \frac{1}{\gamma} \ln |\Theta|$, where $|\Theta|$ denotes the number of feasible channel selection profiles of all APs.

Cross-References

- Auction
- Fairness
- Preferences and Utility Functions
- Sharing Economics

References

- Beresford A, Stajano F (2003) Location privacy in pervasive computing. In: IEEE pervasive computing 2003
- Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press
- Chen X, Huang J (2013) Database-assisted distributed spectrum sharing. IEEE J Sel Areas Commun 31(11):2349–2361
- Chen M, Liew S, Shao Z, Kai C (2010) Markov approximation for combinatorial network optimization. In: INFOCOM, 2010 proceedings IEEE. IEEE, pp 1–9
- Chen X, Gong X, Yang L, Zhang J (2016) Exploiting social tie structure for cooperative wireless networking: a social group utility maximization framework. IEEE/ACM Trans Netw 24(6):3593–3606
- FCC (September 23, 2010) Second memorandum opinion and order
- Freudiger J, Manshaei MH, Hubaux JP, Parkes DC (2009) On non-cooperative location privacy: a game-theoretic analysis. In: ACM CCS 2009
- Gong X, Chen X, Xing K, Shin DH, Zhang M, Zhang J (2017) From social group utility maximization to personalized location privacy in mobile networks. IEEE/ACM Trans Netw 25(3):1703–1716
- Hindriks J, Myles GD (2013) Intermediate public economics. MIT Press, Cambridge
- Kauffmann B, Baccelli F, Chaintreau A, Mhatre V, Papagiannaki K, Diot C (2007) Measurement-based self organization of interfering 802.11 wireless access networks. In: 26th IEEE international conference on computer communications (INFOCOM 2007). IEEE, pp 1451–1459
- Mas-Colell A, Whinston MD, Green JR et al (1995) Microeconomic theory, vol 1. Oxford University Press, New York

- Osborne M, Rubinstein A (1994) A course in game theory. MIT Press, London
- Yang L, Kim H, Zhang J, Chiang M, Tan CW (2013) Pricing-based decentralized spectrum access control in cognitive radio networks. *IEEE/ACM Trans Netw (TON)* 21(2):522–535
- Young H (2001) Individual strategy and social structure: an evolutionary theory of institutions. Princeton University Press, Princeton

Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices

Yi Yuan¹ and Zhiguo Ding²

¹School of Computing and Communications, Lancaster University, Lancaster, UK

²School of Electrical and Electronic Engineering, University of Manchester, Manchester, UK

Synonyms

Beamforming; Cognitive radio; Physical layer security; RF-based energy harvest

Definition

- Cognitive radio (CR) paradigm is an innovative solution to conciliate the existing conflict between the ever-increasing spectrum demand and the inefficient spectrum utilization by enabling a dynamic spectrum access (DSA) in cellular networks.
- Energy harvest (EH)-based radio frequency (RF) is a promising technique to prolong the low-energy devices in the fifth generation (5G) cellular networks.
- Physical layer (PHY) security is a promising mechanism to provide flexible secure communications without the perfect secret key management and distribution.

Historical Background

With the wide usage of wireless communication technologies, wireless devices consume the huge energy to support the escalating data rate and ubiquitous wireless services. However, the huge energy consumption might affect the lifetime of the wireless devices, especially for low-energy devices in cellular networks (Krikidis et al. 2014). Therefore, it is important to find a way to provide enough energy for the wireless devices for supporting the high-quality communication without increasing the dimension of the battery. Since the RF signal is capable to carry information and energy at the same time, in practice (Krikidis et al. 2014; Ding et al. 2015; Lu et al. 2015), energy harvesting at the receivers can be enhanced by increasing the energy of the information-carrying signal. However, the high-powered signals have a higher probability to be eavesdropped due to a higher potential for information leakage. As a consequence, secure communication with energy transfer has been recognized as a crucial issue in cellular networks.

Based on the benefit of PHY security (Wyner 1975), the concept of secure communication in energy harvesting systems has been recently developed in cellular networks (Liu et al. 2014; Shi et al. 2015; Ng et al. 2014; Yuan and Ding 2018). The secrecy performance was investigated based on the assumption that the base station has perfect channel state information (CSI) at the beginning of each scheduling slot (Ng et al. 2014; Shi et al. 2015). With perfect CSI, the base station can design the optimal resource allocation algorithm to efficiently balance the secrecy performance and harvested energy. However, the perfect CSI obtained at the base station reduces spectral efficiency due to a large amount of CSI feedback and channel estimations. Compared to the works assuming the availability of perfect CSI (e.g., Ng et al. 2014; Shi et al. 2015), the works in Ng et al. (2014), Wang and Wang (2015), and Yuan and Ding (2018) propose imperfect CSI available at the base station. In Ng et al. (2014) and Wang and Wang (2015), a deterministic CSI error model was adopted to

investigate the secrecy performance. In addition, the work in Yuan and Ding (2018) investigates the system security by considering a probabilistic CSI error model, which the channel error satisfies the Gaussian distribution.

This entry takes the work in Yuan and Ding (2018) as an example to investigate the secrecy performance of energy transfer systems in the cellular network under the imperfect CSI model. Then, some promising open problems in this area will be discussed.

Throughout this entry, matrices and vectors are denoted by boldface capital letters and lowercase letters, respectively. $\mathbf{A}^H$, $\text{Tr}(\mathbf{A})$, and $|\mathbf{A}|$ represent Hermitian (conjugate) transpose, trace, and determinant of a matrix $\mathbf{A}$, respectively. $\mathbf{A} \succeq \mathbf{0}$ ($\mathbf{A} \succ \mathbf{0}$) indicates that $\mathbf{A}$ is a Hermitian positive semi-definite (definite) matrix. $\text{vec}(\mathbf{A})$ is vectorization of matrix $\mathbf{A}$. $\mathbf{I}$ and $(\cdot)^{-1}$ denote an identity matrix and the inverse of a matrix, respectively. $[\cdot]^+$ denotes $\max\{x, 0\}$ and $\otimes$ is the Kronecker product. $\|\cdot\|_2$ and $\|\cdot\|_F$ represent the Euclidean norm and the Frobenius norm, respectively. $\Re\{\cdot\}$ extracts the real part from a vector. $\mathcal{CN}(\mathbf{u}, \mathbf{\Sigma})$ denotes the distribution of a circularly symmetric complex Gaussian vector with mean vector $\mathbf{u}$ and covariance matrix $\mathbf{\Sigma}$, and $\sim$ means “distributed as”. $\mathbb{C}^{x \times y}$ is the set of complex $x \times y$ matrix, and $\Pr\{\cdot\}$ is the probability of an event.

Key Applications

System Model

Consider an underlay cognitive MIMO scenario which consists of a secondary base station (SBS), $M + 1$ secondary receivers (SU), and K primary receivers (PU). The SBS, the SUs, and the PUs are equipped with N_t antennas, N_s antennas, and N_p antennas, respectively. The channels from the SBS to all the receivers are assumed to undergo quasi-static block fading. It is assumed that SUs are the low-energy devices in cellular networks. Hence, in order to prolong the lifetime of the SUs, it is assumed that only one SU can be acted as the information receivers (IR) to decode information and other M SUs are acted as the energy receiver (ER) to harvest and store energy for

future use. However, if ERs are malicious, they may attempt to eavesdrop the information rather than to harvest energy since they are deployed in the same range of service coverage. To protect the received information by the desired receiver, secure communication should be considered by regarding ERs as the potential eavesdroppers and regarding PUs as the friendly users. The SBS sends the coded confidential signal and artificial noise (AN) signal to all the receivers, denoted by $\mathbf{x} = \mathbf{s} + \mathbf{v}$, where $\mathbf{s} \in \mathbb{C}^{N_t \times 1}$ and $\mathbf{v} \in \mathbb{C}^{N_t \times 1}$. It is assumed that $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q})$ with $\mathbf{Q} \succeq \mathbf{0}$ and $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \mathbf{V})$ with $\mathbf{V} \succeq \mathbf{0}$. The signals received at the IR, the m th ER, and the k th PU can be given, respectively

$$\mathbf{y} = \mathbf{H}^H \mathbf{x} + \mathbf{n}, \quad (1a)$$

$$\mathbf{y}_k = \mathbf{G}_k^H \mathbf{x} + \mathbf{n}_k, k \in K, \quad (1b)$$

$$\mathbf{y}_m = \mathbf{H}_m^H \mathbf{x} + \mathbf{n}_m, m \in M, \quad (1c)$$

where $\mathbf{H} \in \mathbb{C}^{N_t \times N_s}$, $\mathbf{G}_k \in \mathbb{C}^{N_t \times N_p}$, and $\mathbf{H}_m \in \mathbb{C}^{N_t \times N_s}$ are the channel matrices between the SBS and the IR, the k th PU, and the m th ER, respectively. $\mathbf{n}$, $\mathbf{n}_k$, and $\mathbf{n}_m$ are assumed to be independent and identically distributed (i.i.d.) complex Gaussian noise with zero mean and unit variance. Therefore, the maximum achievable secrecy rate between the SBS and the IR can be expressed as (Bloch et al. 2008)

$$R_s = \left[C_{IR} - \max_{m \in M} C_m \right]^+ \quad (2)$$

where $[x]^+$ denotes $\max(0, x)$; $C_{IR} = \log |\mathbf{H}^H(\mathbf{Q} + \mathbf{V})\mathbf{H} + \mathbf{I}| - \log |\mathbf{H}^H\mathbf{V}\mathbf{H} + \mathbf{I}|$; $C_m = \log |\mathbf{H}_m^H(\mathbf{Q} + \mathbf{V})\mathbf{H}_m + \mathbf{I}| - \log |\mathbf{H}_m^H\mathbf{V}\mathbf{H}_m + \mathbf{I}|$. Note that the perfect secrecy rate is achieved when the IR can correctly decode the confidential information at R_s bits per channel use, while all ER can retrieve almost nothing about the secret message. Since the secondary system and the primary system share the same spectrum resource in the underlay CR network, the interference at the k th PU caused by the SBS is given by

$$I_k = \text{Tr}(\mathbf{G}_k^H(\mathbf{Q} + \mathbf{V})\mathbf{G}_k), k \in K. \quad (3)$$

In this entry, the linear energy harvest model is assumed to be applied at each ER. Thus, the harvested energy at the m th ER is given by

$$E_m = \xi_m \text{Tr}(\mathbf{H}_m^H (\mathbf{Q} + \mathbf{V}) \mathbf{H}_m), \quad m \in M \quad (4)$$

where $\xi_m \in (0, 1]$ is the energy conversion efficiency at the m th ER.

Secure Communication Design with Partial Imperfect CSI

In this section, the secrecy performance of the energy-constrained cellular network will be investigated under the partial CSI model which is assumed that perfect CSI for the SBS-to-IR channel and imperfect CSI for the SBS-to-ER channels and the SBS-to-PU channels are available at the SBS during the whole transmission period. To capture the effect caused by imperfect CSI, a probabilistic model will be adopted for modeling the resulting CSI uncertainty.

Probabilistic CSI Error Model

In a probabilistic CSI error model, CSI error is random and follows a certain distribution (Wang et al. 2014). Specifically, CSI error satisfies the complex Gaussian distribution. Accordingly, the actual channel between the SBS and the m th ER as well as the k th PU can be modeled as, respectively

$$\begin{aligned} \mathbf{H}_m &= \hat{\mathbf{H}}_m + \Delta \mathbf{H}_m, \quad m \in M, \\ \mathbf{G}_k &= \hat{\mathbf{G}}_k + \Delta \mathbf{G}_k, \quad k \in K, \end{aligned} \quad (5)$$

where $\hat{\mathbf{H}}_m \in \mathbb{C}^{N_t \times N_s}$ and $\hat{\mathbf{G}}_k \in \mathbb{C}^{N_t \times N_p}$ denote the channel estimation of the m th ER and the k th PU available at the SBS, respectively. $\Delta \mathbf{H}_m$ and $\Delta \mathbf{G}_k$ represent the associated unknown channel uncertainty. Specifically, $\mathbf{e}_{h,m} = \text{vec}(\Delta \mathbf{H}_m) \sim \mathcal{CN}(0, \mathbf{C}_m)$ with $\mathbf{C}_m \succeq \mathbf{0}$ and $\mathbf{e}_{g,k} = \text{vec}(\Delta \mathbf{G}_k) \sim \mathcal{CN}(0, \mathbf{D}_k)$ with $\mathbf{D}_k \succeq \mathbf{0}$.

Efficient Beamforming Design

Efficient beamforming design has been demonstrated to have a significant impact on the security performance under the perfect CSI scenarios

(Ng et al. 2014; Shi et al. 2015). Motivated by this, a robust outage design will be proposed to maximize the secrecy rate in (2) under the transmit power constraint and the outage constraints for EH and interference power. Therefore, an outage-constrained secrecy rate maximization (OC-SRM) problem is formulated as follows:

$$(P1) : \max_{\mathbf{Q}, \mathbf{V}, R} R \quad (6a)$$

$$s.t. \quad \Pr\{\min_{m \in M} \{C_{IR} - C_m\} \geq R\} \geq 1 - \rho, \quad \forall m, \quad (6b)$$

$$\Pr\{\xi \text{Tr}(\mathbf{H}_m^H (\mathbf{Q} + \mathbf{V}) \mathbf{H}_m) \geq \eta_m\} \geq 1 - \rho_m, \quad \forall m, \quad (6c)$$

$$\Pr\{\text{Tr}(\mathbf{G}_k^H (\mathbf{Q} + \mathbf{V}) \mathbf{G}_k) \leq \Gamma_k\} \geq 1 - \rho_k, \quad \forall k, \quad (6d)$$

$$\text{Tr}(\mathbf{Q} + \mathbf{V}) \leq P, \quad \mathbf{Q} \succeq \mathbf{0}, \quad \mathbf{V} \succeq \mathbf{0}, \quad (6e)$$

where η_m and Γ_k denote the minimum required power transfer to the m th ER and the maximum interference power to the k th PU, respectively. P is the maximum transmit power supply at the SBS. ρ , ρ_m , and ρ_k denote the maximum outage probability associated with the secrecy rate, the EH of the m th ER, and the interference power of the k th PU, respectively. Constraints in (6e) guarantee matrices $\mathbf{Q}$ and $\mathbf{V}$ to be positive semi-definite Hermitian matrices. This outage-constrained problem (P1) emphasizes service fidelity. Specifically, under CSI errors, a feasible solution to (P1) guarantees that the IR can still reliably achieve R secrecy message at least $(1 - \rho)$ and the m th ER can harvest η_m energy of the time.

Remark 1 The considered OC-SRM problem may be infeasible when the EH and interference power requirements are stringent and/or the channels are in unfavorable conditions. Without loss of generality, in the rest of this entry, it is assumed that P1 is always feasible for studying the efficient beamforming design.

Problem (P1) is non-convex and very challenging to solve (Boyd and Vandenberghe 2004). The main obstacle to solve P1 lies in the outage

probability constraints (6b), (6c), and (6d), which do not have the tractable closed-form expression. In the following, an efficient algorithm will be proposed to derive the safe approximation to (P1). The “safe” means that the obtained feasible solution always satisfies the outage constraints (6b), (6c), and (6d). According to the challenge 1 proposed in Wang et al. (2014), the idea is to find out the upper bound f of each outage constraint, which satisfies the following condition:

$$\Pr\{\mathbf{e}^H \mathbf{Q} \mathbf{e} + 2\Re\{\mathbf{e}^H \mathbf{r}\} + s < 0\} \leq f(\mathbf{Q}, \mathbf{r}, s), \quad (7)$$

where $\mathbf{e} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. In order to solve the outage secrecy rate constraint in (6b), the following mutual information approximation at the low signal-to-noise ratio (SNR) needs to be considered (Verdú 2002; Wang and Palomar 2009):

$$\log |\mathbf{I} + \frac{1}{\sigma^2} \mathbf{H} \mathbf{Q} \mathbf{H}^H| = \frac{1}{\sigma^2} \text{Tr}\{\mathbf{H} \mathbf{Q} \mathbf{H}^H\} + \mathcal{O}(\frac{1}{\sigma^2} \|\mathbf{H} \mathbf{Q} \mathbf{H}^H\|). \quad (8)$$

Beside this approximation, the following lemma also needs to be used.

Lemma 1 [Jose et al. 2011, Lemma 2:] Define a positive matrix $\mathbf{E} \succ \mathbf{0}$, $\mathbf{E} \in \mathbb{C}^{N \times N}$ with integer N , there exist a function, $v(\mathbf{S}, \mathbf{E}) = -\text{Tr}(\mathbf{S}\mathbf{E}) + \ln |\mathbf{S}| + N$, satisfies the following equation:

$$\ln |\mathbf{E}^{-1}| = \max_{\mathbf{S} \in \mathbb{C}^{N \times N}, \mathbf{S} \succeq \mathbf{0}} v(\mathbf{S}, \mathbf{E}) \quad (9)$$

and the optimal solution to the right-hand side of (9) is $\mathbf{S}^* = \mathbf{E}^{-1}$.

Based on the low SNR approximation and Lemma 1, the secrecy rate expression can be reformulated to the same form in the left-hand side of inequality in (7).

Lemma 2 [Wang et al. 2014, Lemma 1:] Consider the following outage constraint

$$f(\mathbf{A}, \mathbf{u}, c) \Leftrightarrow \Pr\{\mathbf{x}^H \mathbf{A} \mathbf{x} + 2\text{Re}\{\mathbf{x}^H \mathbf{u}\} + c \geq 0\} \geq 1 - \rho, \quad (10)$$

where $\mathbf{A} \in \mathbb{C}^{N \times N}$ is a complex Hermitian matrix, $\mathbf{a} \in \mathbb{C}^{N \times 1}$, $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ and fixed $\rho \in (0, 1]$. With any slack variables λ and μ , the outage constraint can be equivalently transformed into the following three inequalities:

$$\begin{cases} \text{Tr}(\mathbf{A}) - \sqrt{-2 \ln(\rho)} \lambda + \ln(\rho) \mu + c \geq 0, \\ \left\| \begin{bmatrix} \text{vec}(\mathbf{A}) \\ \sqrt{2} \mathbf{u} \end{bmatrix} \right\| \leq \lambda, \\ \mu \mathbf{I}_N + \mathbf{A} \succeq \mathbf{0}, \quad \mu \geq 0. \end{cases} \quad (11)$$

and the optimal solution to the right-hand side of (9) is $\mathbf{S}^* = \mathbf{E}^{-1}$.

By applying Lemma 2 and setting $\mathbf{e}_m = \mathbf{C}_m^{\frac{1}{2}} \mathbf{v}_m$, $\mathbf{e}_k = \mathbf{D}_k^{\frac{1}{2}} \mathbf{x}_k$, all outage constraints in problem (P1) can be converted into the deterministic constraints. Therefore, the safe approximation problem of (P1) is given by

$$\text{(P1-APP)} : \max_{\mathbf{Q}, \mathbf{V}, \mathbf{S}, \mathbf{S}_m, R, \lambda_m, \mu_m, \alpha_m, \beta_m, v_k, \omega_k} R \quad (12a)$$

$$\text{s.t. } \text{Tr}(\mathbf{C}_m^{\frac{1}{2}} \mathbf{\Theta}_m \mathbf{C}_m^{\frac{1}{2}}) + \sqrt{-2 \ln(\rho)} \lambda_m - \ln(\rho) \mu_m + c_m \leq 0, \quad \forall m, \quad (12b)$$

$$\text{Tr}(\mathbf{C}_m^{\frac{1}{2}} \mathbf{\Xi} \mathbf{C}_m^{\frac{1}{2}}) - \sqrt{-2 \ln(\rho_m)} \alpha_m + \ln(\rho_m) \beta_m + \widehat{\mathbf{h}}_m^H \mathbf{\Xi} \widehat{\mathbf{h}}_m - \frac{\eta_m}{\xi} \geq 0, \quad \forall m, \quad (12c)$$

$$\text{Tr}(\mathbf{D}_k^{\frac{1}{2}} \mathbf{\Xi} \mathbf{D}_k^{\frac{1}{2}}) + \sqrt{-2 \ln(\rho_k)} v_k - \ln(\rho_k) \omega_k + \widehat{\mathbf{g}}_k^H \mathbf{\Xi} \widehat{\mathbf{g}}_k - \Gamma_k \leq 0, \quad \forall k, \quad (12d)$$

$$\left\| \begin{bmatrix} \text{vec}(\mathbf{C}_m^{\frac{1}{2}} \mathbf{\Theta}_m \mathbf{C}_m^{\frac{1}{2}}) \\ \sqrt{2} \mathbf{C}_m^{\frac{1}{2}} \mathbf{\Theta}_m \mathbf{h}_m \end{bmatrix} \right\| \leq \lambda_m, \quad \forall m, \quad (12e)$$

$$\left\| \begin{bmatrix} \text{vec}(\mathbf{C}_m^{\frac{1}{2}} \mathbf{\Xi} \mathbf{C}_m^{\frac{1}{2}}) \\ \sqrt{2} \mathbf{C}_m^{\frac{1}{2}} \mathbf{\Xi} \hat{\mathbf{h}}_m \end{bmatrix} \right\| \leq \alpha_m, \quad \forall_m, \quad (13a)$$

$$\left\| \begin{bmatrix} \text{vec}(\mathbf{D}_k^{\frac{1}{2}} \mathbf{\Xi} \mathbf{D}_k^{\frac{1}{2}}) \\ \sqrt{2} \mathbf{C}_m^{\frac{1}{2}} \mathbf{\Xi} \hat{\mathbf{g}}_k \end{bmatrix} \right\| \leq \nu_k, \quad \forall_k, \quad (13b)$$

$$\mu_m \mathbf{I}_{N_t \times N_s} - \mathbf{C}_m^{\frac{1}{2}} \mathbf{\Theta}_m \mathbf{C}_m^{\frac{1}{2}} \succeq \mathbf{0}, \quad \mu_m \geq 0, \quad \forall_m, \quad (13c)$$

$$\beta_m \mathbf{I}_{N_t \times N_s} + \mathbf{C}_m^{\frac{1}{2}} \mathbf{\Xi} \mathbf{C}_m^{\frac{1}{2}} \succeq \mathbf{0}, \quad \beta_m \geq 0, \quad \forall_m, \quad (13d)$$

$$\omega_k \mathbf{I}_{N_t \times N_p} - \mathbf{D}_k^{\frac{1}{2}} \mathbf{\Xi} \mathbf{D}_k^{\frac{1}{2}} \succeq \mathbf{0}, \quad \omega_k \geq 0, \quad \forall_k, \quad (13e)$$

$$\mathbf{S} \succeq \mathbf{0}, \mathbf{S}_m \succeq \mathbf{0}, \quad (6e), \quad (13f)$$

E

where $\hat{\mathbf{h}}_m = \text{vec}(\hat{\mathbf{H}}_m)$, $\hat{\mathbf{g}}_k = \text{vec}(\hat{\mathbf{G}}_k)$, $\mathbf{\Theta}_m = (\mathbf{S}_m^T \otimes (\mathbf{Q} + \mathbf{V}) - \mathbf{I}_{N_t} \otimes \mathbf{V})$, $c_m = \hat{\mathbf{h}}_m^H \mathbf{\Theta}_m \hat{\mathbf{h}}_m - \text{Tr}(\mathbf{H}^H (\mathbf{Q} + \mathbf{V}) \mathbf{H}) + \text{Tr}(\mathbf{S}(\mathbf{H}^H \mathbf{V} \mathbf{H} + \mathbf{I})) - \ln |\mathbf{S}| - 2N_s + \text{Tr}(\mathbf{S}_m) - \ln |\mathbf{S}_m| + R$, and $\mathbf{\Xi} = \mathbf{I} \otimes (\mathbf{Q} + \mathbf{V})$.

Note that problem (P1-APP) is non-convex w.r.t. all the variables jointly (Boyd and Vandenberghe 2004). However, (P1-APP) becomes convex when fixing either $(\mathbf{Q}, \mathbf{V})$ or $(\mathbf{S}, \mathbf{S}_m)$. Hence, an alternating optimization (AO) algorithm is proposed to optimize all variables in (P1-APP) (Schubert and Boche 2004). Specifically, problem (P1-APP) becomes a convex problem by fixing $(\mathbf{S}, \mathbf{S}_m)$. Then, the same operation can be considered to optimize $(\mathbf{S}, \mathbf{S}_m)$. The summarization of the proposed AO algorithm is presented in the following:

-
- 1 Initialize
 $n = 1, \varepsilon > 0, \mathbf{S}^0 = \mathbf{I}, \mathbf{S}_m^0 = \mathbf{I}, m = 1, \dots, M$.
 - 2 At the l th iteration
 - a) Solve (P1-APP) with fixed $\mathbf{S} = \mathbf{S}^{l-1}$ and $\mathbf{S}_m = \mathbf{S}_m^{l-1}$ to obtain $(\mathbf{Q}^{l-1}, \mathbf{V}^{l-1})$.
 - b) Solve (P1-APP) with fixed $\mathbf{Q} = \mathbf{Q}^{l-1}$ and $\mathbf{V} = \mathbf{V}^{l-1}$ to obtain $(\mathbf{S}^l, \mathbf{S}_m^l, R^l)$.
 - 3 Update $l = l + 1$ and repeat step (2) until the optimal value is found.
-

It notes that the convex problems in step a) and b) can be efficiently solved by using off-the-shelf solver, e.g., CVX (Grant and Boyd 2011). Moreover, the optimal solutions generated at the $l - 1$ th iteration are also the

feasible values at the l th iteration. For this reason, the proposed AO algorithm generates a nondecreasing sequence of the secrecy rate and converges to a stationary point. In addition, the smaller predefined accuracy is, the more iterative steps are need. Hence, the time of solving the proposed algorithm can be roughly estimated as the time of solving the two subproblems multiplies the number of iterations.

Secure Communication Design with Full Imperfect CSI

Although the transmitter can obtain perfect CSI of the user through measuring the uplink pilots in the handshaking/beacon signals via channel reciprocity, CSI of the user might not be perfectly obtained in some special communication scenarios. Therefore, in this section, a more general channel model will be introduced in which it is assumed that imperfect CSI of the SBS-to-IR channel are available at the SBS during the whole transmission period.

Probabilistic CSI Error Model

In the above section, the error model of the SBS-to- m th ER link and the SBS-to- k th PU link has been derived. Therefore, the actual channel between the SBS and the IR can be modeled as

$$\mathbf{H} = \hat{\mathbf{H}} + \Delta \mathbf{E}, \quad (14)$$

where $\hat{\mathbf{H}} \in \mathbb{C}^{N_t \times N_s}$ is the channel estimate of the IR; $\Delta \mathbf{E}$ denote the associated unknown channel uncertainty; and $\mathbf{e} = \text{vec}(\Delta \mathbf{E}) \sim \mathcal{CN}(0, \mathbf{Z})$ with $\mathbf{Z} \succeq \mathbf{0}$.

Efficient Beamforming Design

In this subsection, the efficient beamforming is designed based on the practical channel uncertainty model. In this case, the OC-SRM problem is the same to (P1), but the secrecy rate expression in (2) needs to be carefully reformulated since channel errors are included in both rate expressions. Based on the low SNR approximation in (8) and Lemma 1, the secrecy outage constraint (6b) can be approximated to the following constraint by setting $\mathbf{e} = \mathbf{Z}^{\frac{1}{2}} \mathbf{v}$:

$$\begin{aligned} & \Pr \left\{ [\mathbf{v}^H \mathbf{v}_m^H] \begin{bmatrix} \mathbf{Z}^{\frac{1}{2}} \boldsymbol{\varpi} \mathbf{Z}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_m^{\frac{1}{2}} \boldsymbol{\Psi}_m \mathbf{C}_m^{\frac{1}{2}} \end{bmatrix} [\mathbf{v}^H \mathbf{v}_m^H]^H \right. \\ & + 2\Re \left([\mathbf{v}^H \mathbf{v}_m^H] \begin{bmatrix} \mathbf{Z}^{\frac{1}{2}} \boldsymbol{\varpi} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_m^{\frac{1}{2}} \boldsymbol{\Psi}_m \end{bmatrix} [\hat{\mathbf{h}}^H \hat{\mathbf{h}}_m^H]^H \right) \\ & \left. + [\hat{\mathbf{h}}^H \hat{\mathbf{h}}_m^H] \begin{bmatrix} \boldsymbol{\varpi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_m \end{bmatrix} [\hat{\mathbf{h}}^H \hat{\mathbf{h}}_m^H]^H - c_m \leq 0 \right\} \geq 1 - \rho, \end{aligned} \quad (15)$$

where $\hat{\mathbf{h}} = \text{vec}(\hat{\mathbf{H}})$, $\boldsymbol{\varpi} = \mathbf{S}^T \otimes \mathbf{V} - \mathbf{I}_{N_t} \otimes (\mathbf{Q} + \mathbf{V})$, $\boldsymbol{\Psi}_m = \mathbf{S}_m^T \otimes (\mathbf{Q} + \mathbf{V}) - \mathbf{I}_{N_t} \otimes \mathbf{V}$, and $c_m = -\text{Tr}(\mathbf{S}) + \ln |\mathbf{S}| - \text{Tr}(\mathbf{S}_m) + \ln |\mathbf{S}_m| - R$. The outage constraint has the similar form with (10); therefore, the outage constraint (15) can be equivalently converted to the following deterministic constraint by applying Lemma 2:

$$\begin{cases} \text{Tr}(\mathbf{A}_{s,m}) + \sqrt{-2 \ln(\rho) \theta_{s,m} - \ln(\rho) \delta_{d,m} + \sigma_{s,m}} \leq 0, \\ \left\| \begin{bmatrix} \text{vec}(\mathbf{A}_{s,m}) \\ \sqrt{2} \mathbf{B}_{s,m} \end{bmatrix} \right\| \leq \theta_{s,m}, \\ \delta_{s,m} \mathbf{I}_{N_t \times N_s} - \mathbf{A}_{s,m} \succeq \mathbf{0}, \quad \delta_{s,m} \geq 0, \end{cases} \quad (16)$$

$$\begin{aligned} \text{where } \mathbf{A}_{s,m} &= \begin{bmatrix} \mathbf{Z}^{\frac{1}{2}} \boldsymbol{\varpi} \mathbf{Z}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_m^{\frac{1}{2}} \boldsymbol{\Psi}_m \mathbf{C}_m^{\frac{1}{2}} \end{bmatrix}, \\ \mathbf{B}_{s,m} &= \begin{bmatrix} \mathbf{Z}^{\frac{1}{2}} \boldsymbol{\varpi} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_m^{\frac{1}{2}} \boldsymbol{\Psi}_m \end{bmatrix} [\hat{\mathbf{h}}^H \hat{\mathbf{h}}_m^H]^H, \quad \sigma_{s,m} = \end{aligned}$$

$[\hat{\mathbf{h}}^H \hat{\mathbf{h}}_m^H] \begin{bmatrix} \boldsymbol{\varpi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_m \end{bmatrix} [\hat{\mathbf{h}}^H \hat{\mathbf{h}}_m^H]^H - c_m$, $\mathbf{v}_{s,m} = [\mathbf{v}^H \mathbf{v}_m^H]^H$. The EH and interference power outage constraints based on the full uncertainty model are the same as that of partial model. Hence, the OC-SRM problem based on the full uncertainty model is reformulated by

$$\begin{aligned} \text{(P2): } & \max_{\mathbf{Q}, \mathbf{V}, \mathbf{S}, \mathbf{S}_m, R, \delta_{s,m}, \theta_{s,m}, \alpha_m, \beta_m, \nu_k, \omega_k} R \\ \text{s.t. } & \text{(12c), (12d), (13a), (13b), (13d), (13f), (16), (17)} \end{aligned}$$

Similar to (P1-APP), problem (P2) is non-convex due to the coupled variables. Therefore, the AO algorithm proposed for solving (P1-APP) also can be used to solve (P2).

Numerical Results

In this section, the system performance of the considered security cellular network under the imperfect CSI model will be evaluated via computer simulation results. All estimated channels are randomly generated following an i.i.d. complex Gaussian distribution with zero mean and unit variance. The number of antenna at the SBS sets to 4, $N_t = 4$, and at the receivers sets to 3, $N_s = N_p = 3$. In addition, the number of the ER and the PU is set to 3, $M = K = 3$. The EH requirement and the interference power tolerance are considered as -5 and 5 dBm, respectively. The maximum outage is $\rho_s = \rho_m = \rho_k = \rho = 0.01$, $\forall_m$, $\forall_k$, and the error variance is $\mathbf{Z} = \mathbf{C}_m = \mathbf{D}_k = \sigma^2 \mathbf{I}$, $\forall_m$, $\forall_k$.

Benchmark Schemes

- (1) *Perfect CSI Scheme*: The perfect CSI scheme is taken as one benchmark to provide an upper-bound secrecy performance. In this scheme, the SBS knows the exact CSI of all links.
- (2) *Non-robust Scheme*: To show the benefits of the proposed robust design, the secrecy performance of the non-robust design is taken as the second benchmark scheme, in which the SBS does not spend the additional transmit power to overcome the

effects caused by the channel errors. In this case, the actual channels included error that can be randomly generated based on their distribution, i.e., $\text{vec}(\mathbf{H}) \sim \mathcal{CN}(\text{vec}(\hat{\mathbf{H}}), \mathbf{Z})$, $\text{vec}(\mathbf{H}_m) \sim \mathcal{CN}(\text{vec}(\hat{\mathbf{H}}_m), \mathbf{C}_m)$, $\forall m$, $\text{vec}(\mathbf{G}_k) \sim \mathcal{CN}(\text{vec}(\hat{\mathbf{G}}_k), \mathbf{D}_k)$, $\forall k$.

- (3) *Worst-Case Scheme*: To show the benefits of the proposed outage design, the worst-case design is taken as the third benchmark scheme, in which the deterministic CSI uncertainty model is considered to capture the effects caused by the channel errors. In order to have a fair comparison with the proposed outage design scheme, the error bound used in the worst-case design is generated by $\varepsilon = \sqrt{\frac{\sigma^2 F_{2N_t \times N}^{-1}(1-\rho)}{2}}$, where $F_{2N_t \times N}^{-1}(\cdot)$ denotes the inverse cumulative distribution function (CDF) of the chi-square distribution with $2N_t \times N$ degrees of freedom (Wang et al. 2014).

Secrecy Performance

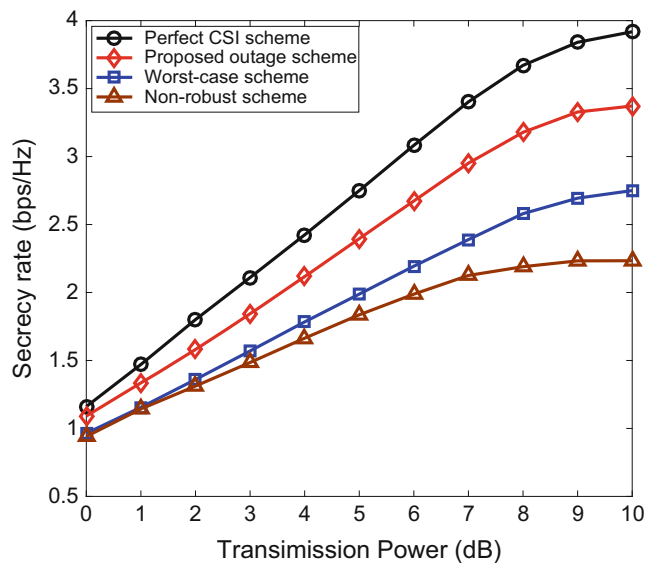
The secrecy performance of the energy-constrained cellular network is shown under the partial imperfect CSI model. In Figs. 1 and 2, the secrecy performance of the proposed outage scheme is compared with the benchmark

schemes defined previously under the different transmit power P and error variance σ^2 . From these figures, one can observe that the proposed outage scheme with the efficient beamforming design outperforms the comparable schemes, i.e., non-robust scheme and worst-case scheme. As shown in Fig. 1, the secrecy rates of all the schemes increase with respect to the transmit power P and finally become saturate, which is due to the interference power constraint to the PU. Moreover, the performance generated by the proposed outage design is closer to the perfect CSI scheme compared with the other two schemes, so the proposed outage design can significantly improve the secrecy performance for the imperfect CSI models. From Fig. 2, the secrecy performance of all the schemes is affected by the channel estimation error. The proposed outage design outperforms the other two benchmark schemes, and the performance gap becomes large as error variance increases. Moreover, the secrecy performance of strict interference power tolerance ($I = -6\text{ dB}$) is worse than that of the relaxed tolerance ($I = 2\text{ dB}$).

The secrecy performance of the energy-constrained cellular network is focused based on the full channel uncertainty model. In Figs. 3

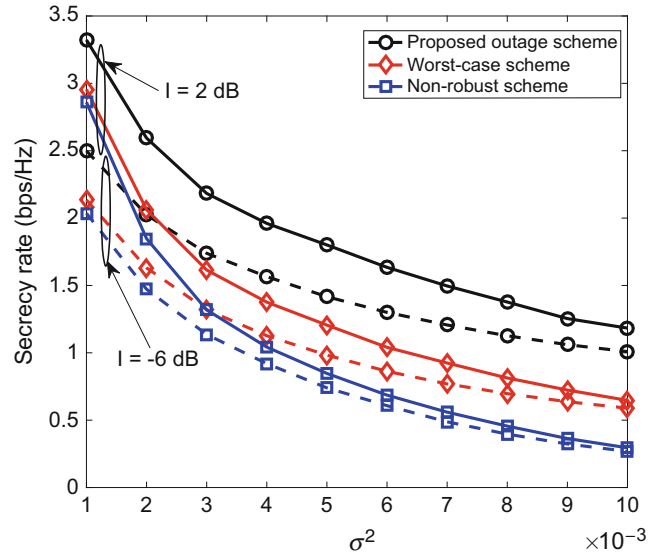
Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices, Fig. 1

The secrecy rate of the different methods versus the transmit power based on the partial uncertainty model with $I = -5\text{ dB}$, $\sigma^2 = 0.001$, and $\eta_m = \eta = 5\text{ dBm}$



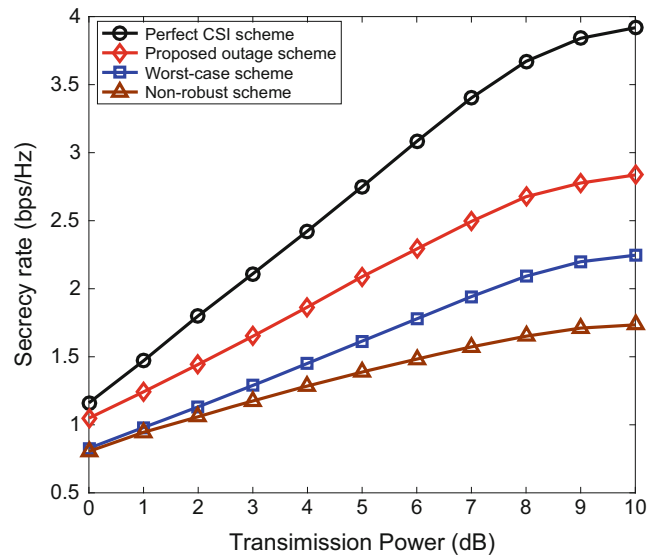
Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices, Fig. 2

The secrecy rate versus error variance σ^2 for the different interference power based on the PCU model with $P = 5$ dB, $\eta_m = \eta = 5$ dBm



Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices, Fig. 3

The secrecy rate of the different methods versus the transmit power based on the full uncertainty model with $I = -5$ dB, $\sigma^2 = 0.001$, $\eta_m = \eta = 5$ dBm

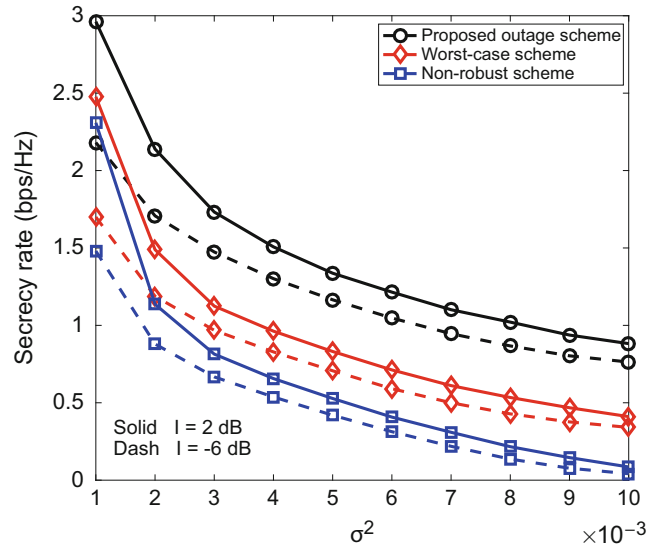


and 4, the secrecy performance of the proposed outage scheme is compared with the benchmark schemes under the different transmit power P and error variance σ^2 . As shown in Fig. 3, the secrecy performance of all the schemes has a significant improvement with the sufficient transmit power supply, especially for the outage scheme. This implies that the proposed outage scheme is capable to generate the efficient

beamforming to enhance the system security. The results in Fig. 4 indicate that the secrecy rates of all the schemes have the same degradation trend but the proposed outage scheme still can achieve the better performance compared with the other two schemes. The reason is that the outage scheme can efficiently utilize the transmit power to overcome the error estimation effects.

Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices, Fig. 4

The secrecy rate versus error variance σ^2 for the different interference power based on the full uncertainty model with $P = 5$ dB, $\eta_m = \eta = 5$ dBm



Conclusion

This entry has provided a fundamental analysis for investigating the secrecy performance of the underlay energy-constrained cellular network with imperfect CSI. To capture the effects of the channel uncertainty, a probabilistic CSI error model was proposed. Then, based on the proposed channel uncertainty models, the OR-SRM problem was formulated to design the efficient resource allocation in order to improve the system security. The proposed optimization problem was a non-convex problem due to the outage constraints. By applying the transformation and approximation, the non-convex problem was decoupled to two conic convex problems. Then, an AO algorithm was proposed to iteratively solve these two convex problems. The considered outage design was demonstrated to outperform the worst-case and non-robust design in terms of the secrecy rate via numerical results under the imperfect CSI model in the cellular network.

Future work of interest includes extending the considered outage design to multiple IRs to evaluate the secrecy rate and harvested energy. Furthermore, combining PHY security techniques with cryptography techniques in the higher communication protocol layers to further improve the secrecy performance in energy-constrained

cellular networks is thus another interesting problem to investigate in the future. Finally, secrecy energy efficiency (SEE) and secrecy spectral efficiency (SSE) could be studied to balance the security and energy consumption.

Cross-References

- Cognitive Radio Networks
- Energy Efficiency of Cellular Networks
- Massive MIMO
- Wireless Data Security

References

- Bloch M, Barros J, Rodrigues MRD, McLaughlin SW (2008) Wireless information-theoretic security. *IEEE Trans Inf Theory* 54(6):2515–2534
- Boyd S, Vandenberghe L (2004) *Convex optimization*. Cambridge University Press, New York
- Ding Z, Zhong C, Ng DWK, Peng M, Suraweera HA, Schober R, Poor HV (2015) Application of smart antenna technologies in simultaneous wireless information and power transfer. *IEEE Commun Mag* 53(4):86–93
- Grant M, Boyd S (2011) CVX: matlab software for disciplined convex programming, version 1.21. April 2011 [Online]. Available: <http://cvxr.com/cvx/>
- Jose J, Prasad N, Khojastepour M, Rangarajan S (2011) On robust weighted-sum rate maximization in MIMO

interference networks. In: Proceeding of IEEE ICC, Kyoto

- Krikidis I, Timotheou S, Nikolaou S, Zheng G, Ng DWK, Schober R (2014) Simultaneous wireless information and power transfer in modern communication systems. *IEEE Commun Mag* 52(11):104–110
- Liu L, Zhang R, Chua KC (2014) Secrecy wireless information and power transfer with miso beamforming. *IEEE Trans Signal Process* 62(7):1850–1863
- Lu X, Wang P, Niyato D, Kim DI, Han Z (2015) Wireless networks with rf energy harvesting: a contemporary survey. *IEEE Commun Surv Tuts* 17(2):757–789
- Ng DWK, Lo ES, Schober R (2014) Robust beamforming for secure communication in systems with wireless information and power transfer. *IEEE Trans Wirel Commun* 13(8):4599–4615
- Schubert M, Boche H (2004) Solution of the multiuser downlink beamforming problem with individual sinr constraints. *IEEE Trans Veh Technol* 53(1):18–28
- Shi Q, Xu W, Wu J, Song E, Wang Y (2015) Secure beamforming for mimo broadcasting with wireless information and power transfer. *IEEE Trans Wirel Commun* 14(5):2841–2853
- Verdú S (2002) Spectral efficiency in the wideband regime. *IEEE Trans Inf Theory* 48(6):1319–1343
- Wang JH, Palomar DP (2009) Worst-case robust MIMO transmission with imperfect channel knowledge. *IEEE Trans Signal Process* 57(8):3086–3100
- Wang S, Wang B (2015) Robust secure transmit design in mimo channels with simultaneous wireless information and power transfer. *IEEE Signal Process Lett* 22(11):2147–2151
- Wang KY, So AMC, Chang TH, Ma WK, Chi CY (2014) Outage constrained robust transmit optimization for multiuser MISO downlinks: tractable approximations by conic optimization. *IEEE Trans Signal Process* 62(21):5690–5705
- Wyner AD (1975) The wire-tap channel. *Bell Syst Tech J* 54(8):1355–1387
- Yuan Y, Ding Z (2018) Outage constrained secrecy rate maximization design with SWIPT in MIMO-CR System. *IEEE Trans Veh Technol* 67(6):5475–5480

Elastic Wireless Resource Management

- [Resource Scheduling in Virtualized Wireless Networks](#)

Electromagnetic Nanonetworks

- [Nanoscale Terahertz Communications](#)

Emergency Medical Services

- [Neighborhood-Based ICT-Driven Support \(NIS\) System of Emergency Medical Services \(EMS\) for Elderly People in Hyper-Aged Societies](#)

Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges

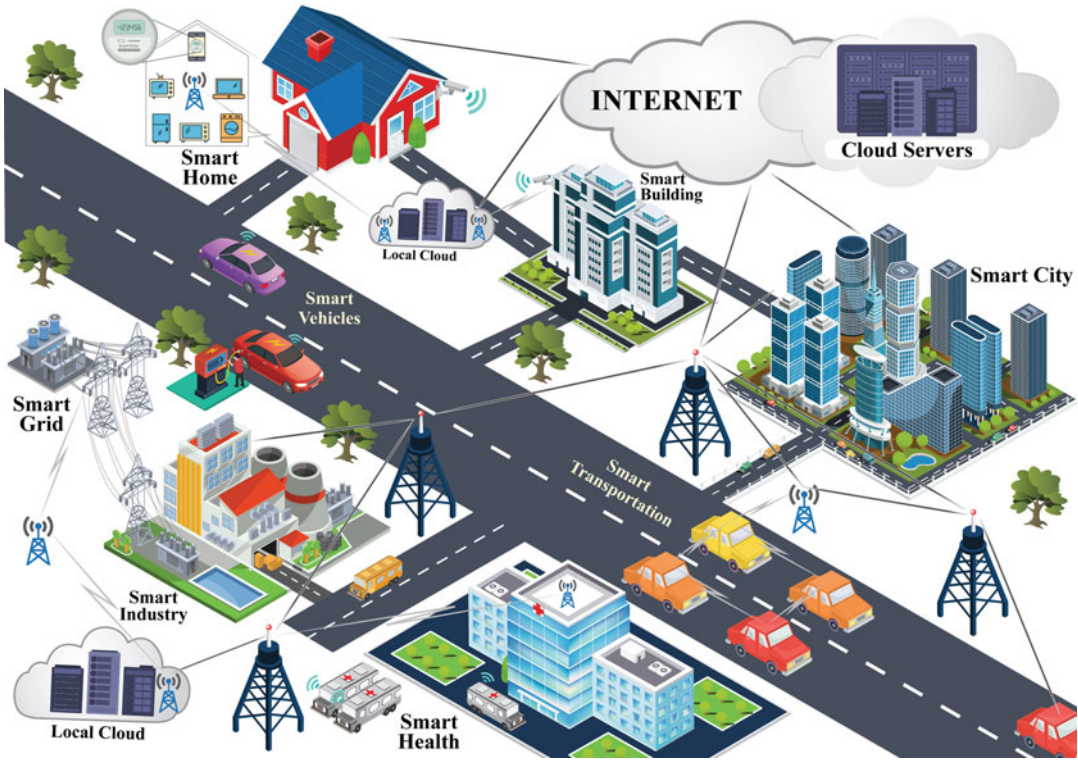
Muhammad Baqer Mollah¹, Sherali Zeadally², and Md. Abul Kalam Azad¹

¹Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

²College of Communication and Information, University of Kentucky, Lexington, KY, USA

Introduction

The Internet of Things (IoT) is a novel paradigm which enables Internet connectivity to all kinds of objects and people (Qiu et al. 2018). Such objects, also called IoT devices, can be sensors, actuators, embedded systems, smart devices, and mobile phones. In this new paradigm, everything and everyone with a common interest can be connected, and for that reason, IoT is often called the Internet of Everything. In IoT systems, IoT devices collect data from physical systems, communicate with gateways for data aggregation, and connect to the Internet to forward the data to the cloud or edge computing devices for further processing and analysis. By connecting IoT devices to the Internet, the IoT ecosystem promises to bring improvements to our quality of lives, environment and system performance in the home, building, city, electric power grid, vehicles, transportation, logistics, healthcare, and many more. Figure 1 illustrates some of emerging IoT application scenarios. Although wireless sensor networks (WSN) are one of the main



E

Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges, Fig. 1
Emerging IoT application scenarios

components of IoT systems, unlike WSN devices, IoT devices are smart enough to take optimal decisions with or without minimal human intervention. According to the recent Cisco report (2018), the number of IoT devices connected to the Internet will reach 50 billion by 2020.

Different communication technologies (wired, optical, and wireless) can be used for data communications among a huge number of IoT devices and for backhaul network scenarios. Wireless technologies will be the best option to connect such IoT devices because of its benefits over wired technologies including easier installation, lower cost infrastructures, mobility support, scalability, and ease of connection. Consequently, in this work, we only focus on wireless technologies. However, traditional popular wireless technologies such as classic Bluetooth, Wi-Fi, LTE, and so on cannot adequately support the envisioned IoT communication networks because of the following reasons:

- (i) They do not have adequate power saving mechanisms to support resource-limited IoT devices.
- (ii) They are mainly designed for human-operated devices.
- (iii) They cannot support a large number of connecting devices through a single base stations or an access point.
- (iv) The frequency bands they use are not suitable for low data rate and long-range communications.
- (v) The use unsuitable wide channel bandwidth that does not support low data rate massive number of devices.

To address the above issues, many researchers, industry groups, and standardization bodies are working towards developing wireless technologies dedicated to IoT and some have already been introduced, often referred to as low power wireless technologies. Hence, it is expected that many

IoT devices would be connected through these technologies to bring people and the physical world onto the Internet.

IoT applications have diverse requirements. This means that more than one wireless technology is necessary to support these applications. Future IoT scenarios will leverage a hybrid network consisting of multiple wireless technologies. Therefore, it is necessary to define the communication requirements of emerging IoT applications so that we can determine the optimal wireless technologies that can fulfill the future IoT vision.

We summarize the main contributions of this article as follows:

- We present the communication requirements of emerging IoT applications.
- We identify wireless technologies and analyze their suitability in meeting the various IoT requirements.
- We present open issues about new research areas needed to improve wireless technologies.

The rest of this article is organized as follows. The next section presents the main communications requirements of IoT scenarios, focusing on a number of specific IoT applications. In section “[Candidate Wireless Technologies for IoT Applications](#),” we present various wireless technologies along with their characteristics. Next, we highlight the pros and cons of each wireless technology when used in IoT applications. In section “[Open Issues](#),” we the open issues to explore future research opportunities and we finally make some concluding remarks in section “[Conclusion](#).”

Communication Requirements of IoT Applications

IoT is paving the way for the emergence of a wide range of applications. In this section, we select some emerging IoT applications and discuss their

communication requirements for various IoT scenarios. IoT applications typically have following communication requirements.

- (i) *Communication latency*: The delay incurred by a particular application during data communication is called communication latency. It directly affects the usability and efficiency of IoT applications. Although latency requirements vary from one application to another, most of the IoT applications are sensitive to latency. Hence, low latency is desirable.
- (ii) *Power consumption*: In general, wireless IoT devices are powered from batteries or energy harvesting. Therefore, the power consumption for wireless IoT devices should be as low as possible.
- (iii) *Data rate*: In many IoT scenarios, most applications require low data rate although some may need a higher data rate. Usually, high data rates will depend on the network bandwidth of the underlying communication channels.
- (iv) *Range*: Wireless technologies need to operate over an adequate range to support IoT applications efficiently and they often extend their coverage using routers, gateway, or access points (AP). The range varies according to the nature of applications.
- (v) *Support for a large number of connections*: The wireless technologies in an IoT ecosystem have to support a large number IoT devices, smart devices, and people from hundreds to billions. Typically, the number of IoT devices is much higher than smartphones.
- (vi) *Mobility*: IoT devices can be either stationary or mobile. For example, there are no mobile IoT devices in smart grid applications, while smart transportation has many mobile devices.

Table 1 presents a summary of the communication requirements of IoT applications. Depending on the particular application scenario, the requirements may differ.

Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges, Table 1
Communication requirements for IoT applications

IoT application	Tolerable latency	Power consumption	Data rate	No. of devices	Mobility	Range
Smart Home	Medium	Low	Low	Low	Low	Short
Smart Building	Medium	Low	Low	Medium	Low	Short–medium
Smart Grid	Medium–high	Low	Low	High	Low	Long
Smart Meter	Medium–high	Low	Low	Low	No	Short
Smart Security	Low	Low	High	Medium	No	Medium
Smart Industry	Medium	Low	Variable	High	Medium	Medium
Smart Health	Low	Low	Low–medium	Medium	Medium	Short–medium
Smart Vehicle	Low	Low	Very high	Medium	Very high	Short
Smart Transportation	Low	Low	Very high	High	Very high	Medium–long
Smart City	Variable	Low	Variable	Variable	Medium	Long

Candidate Wireless Technologies for IoT Applications

In recent years, several wireless technologies have emerged in order to support many of the communication requirements of IoT applications. To choose an appropriate technology for IoT applications, we need to carefully consider the main characteristics of those technologies. In this section, we present the main wireless technologies that can be deployed for IoT applications.

Bluetooth Low Energy (BLE)

BLE is developed by Bluetooth Special Interest Group (SIG) for low power and short range applications (Harris et al. 2016). Compared with classic Bluetooth, its coverage area is 10 times higher, it reduces the communication latency by 15 times, allows multi-hop scenarios, and can be operated by transmitting power that ranges from 0.01 mW to 10 mW. These characteristics make it possible to be a potential candidate for some IoT applications. At the physical layer, BLE uses 40 channels with 2 MHz channel spacing. Among 40 channels, 3 are dedicated. One is used as an advertising channel to discover the devices. Another one is used for setting up a connection and the third one is used for broadcast transmissions. The remaining channels are used as data channels. The dedicated advertising

channels were selected to avoid interference with other wireless technologies which also operate in the 2.4 GHz band. However, the latest BLE specification to date is Bluetooth 5 after versions 4.0, 4.1 and 4.2 (<https://www.bluetooth.com/specifications/bluetooth-core-specification/bluetooth5>). The new features extend the communication range and data rate and simplify the deployment of a BLE network in IoT applications.

WiFi

WiFi becomes an everyday tool already for Internet access in our everyday lives. The WiFi standard IEEE 802.11 was first released in 1997 and many amendments were approved after that. But most of the amendments were optimized and designed for smartphones, tablets, and laptops, not for IoT devices. However, some recently approved IEEE 802.11 amendments have been designed to support different IoT applications which we describe below.

Low Power WiFi (LPW): LPW was introduced to address weaknesses such as the bandwidth, coverage range, and number of connections between traditional networks and the emerging demand for IoT applications. IEEE 802.11ah is the standard for LPW which is an amendment of IEEE 802.11 standard (Park 2015). It designs by modifying the PHY and MAC layers including new short frame formats, advanced

channel access mechanisms, and novel power management mechanisms in order to allow connectivity among IoT devices with WiFi. A new channel access mechanism called Restricted Access Window (RAW) has been introduced to minimize collision probability. Additionally, WiFi Direct is another step in WiFi's evolution that enables direct communications among devices without APs.

WhiteFi: WhiteFi introduces the use of WiFi technology within the TV white space (TVWS) spectrum. The IEEE 802.11af standard describes the use of WiFi on TVWS operation and implementation (Hassan et al. 2017). TVWS is the unused spectrum resource that is basically allocated to TV broadcasters (called primary users) that can be utilized by unlicensed devices (also called secondary users).

DSRC: IEEE 802.11p is a popular standard for dedicated short-range communications (DSRC) (Bazzi et al. 2017; Ligo et al. 2018). It is mainly designed for vehicular communications. The US Federal Communications Commission (FCC) allocates dedicated 75 MHz bandwidth at 5.9 GHz for DSRC. In addition, this bandwidth is divided into seven channels that can be used by safety and nonsafety services.

mmWave: The IEEE 802.11ad working group introduced the operation of multigigabit wireless data transfers over the unlicensed 60 GHz band (i.e., the mmWave band) (Yilmaz et al. 2016). The 60 GHz bands provide 7 GHz bandwidth that can provide up to 6.76 Gbps which can support IoT applications where a very high data rate is required. The PHY layer of this standard has four transmission modes, namely, (i) control mode for beaconing and initial network setup, (ii) single carrier (SC) mode with the data rate up to 4.62 Gbps, (iii) orthogonal frequency division multiplexing (OFDM) mode with the highest data rate of 6.75 Gbps, and (iv) low-power SC mode to support to resource-limited devices.

LoRa

LoRa (Long Range) is one of the emerging wide area wireless technologies (Raza et al. 2017; Navarro-Ortiz et al. 2018) which was initially developed by Semtech

(<http://www.semtech.com/wireless-rf/internet-of-things/what-is-lora/>) and later was promoted by the LoRa Alliance (www.lora-alliance.org/technology). Optimized for low power consumption, LoRa utilizes chirp spread spectrum (CSS) modulation technology to support millions of devices and long-range communication of small data packets. The end IoT devices communicate with gateways through a single hop LoRa wide area network and all gateways are connected to a network server by backhaul connection. The main functions of the network server include filtering duplicate packets from gateways, ensuring security, sending acknowledgments to the gateways, and transmitting the packets to their destinations.

Long Term Evolution (LTE)

LTE and its revision LTE-Advanced (LTE-A) are wide area cellular communications standard developed by third Generation Partnership Project (3GPP) for high-performance communications involving human-operated mobile or similar devices. According to a recent Ericsson Mobility Report (www.ericsson.com/en/mobility-report), in 2017, LTE systems provided communication services to over 2.1 billion of users across the globe. The widespread use of LTE is mainly due to its global standard infrastructure and more flexible radio resource management techniques than other previous cellular technologies. However, the evolution of LTE system promises to meet requirements in large-scale IoT deployments by addressing IoT use cases. Already 3GPP has launched initiatives to modify and optimize the LTE standard to adapt it to IoT applications. These initiatives include LTE for machine type communications (LTE-M) and narrowband IoT (NB-IoT).

LTE-M: LTE-M is a simplified term of LTE-MTC which provides wireless access to IoT applications by reusing LTE installed base stations (Hoymann et al. 2016; Lauridsen et al. 2017). Here, MTC refers to machine type communications where IoT devices are allowed to communicate with each other without human intervention. Today, MTC is an essential part of IoT as it supports devices that access

and communicate with a gateway or access point in various IoT scenarios. Moreover, LTE-M offers device to device (D2D) communications which allow devices in close proximity to communicate directly without going through a base station. This direct connectivity can significantly increase network spectral efficiency, reduce the transmission delay, offload traffic, and avoid congestion in the core network.

NB-IoT: Another initiative by 3GPP called NB-IoT aims to develop a new radio access network to support IoT applications with low data rate (Miao et al. 2018; Xu et al. 2018). Although NB-IoT operates in the narrow band (a new air interface derived from the legacy LTE), NB-IoT inherits the basic features of the LTE standard. However, many features of LTE such as handover mechanism, carrier aggregation, and dual connectivity have been removed to minimize device costs and power consumption.

Table 2 presents a summary of the main characteristics of the wireless technologies we have discussed above.

Pros and Cons of Candidates and Suitable IoT Applications

This section identifies and evaluates viable wireless technologies that can provide efficiently for the various IoT applications shown in Table 1.

To support short-range communication, BLE is promising for many IoT applications mainly due to its low power consumption. Moreover, we foresee an increasing number of smartphones as well as consumer devices equipped with BLE interfaces in the near future. In fact, the BLE interface is already available in most smartphones, and consequently, BLE will have a larger market than ZigBee. The Bluetooth SIG predicts (www.bluetooth.com/pages/mobile-telephony-market.aspx) that by 2018 more than 90% of Bluetooth-enabled smartphones will include BLE. As a result, BLE will become the most common communication medium in IoT scenarios where user to device interactions are required such as in smart home and smart healthcare applications. Regarding interference

issue with 2.4 GHz WiFi, BLE can provide good performance as it uses an adaptive frequency hopping scheme.

The mmWave is well suited for IoT applications, which require high network bandwidth such as high definition video sharing within vehicular communication scenarios. But the main shortcomings of mmWave are its high power consumption, severe propagation loss, and signal attenuation which result in a short-range coverage. However, as the wavelength of 60 GHz is at the millimeter level, it is possible to apply beamforming technique in mmWave by packing a large number of antennas into a small space. Beamforming is a signal processing technique that concentrates power in one direction in order to achieve one directional signal transmission or reception by a smart antenna array. Although mmWave covers about 10 m in omnidirectional broadcast mode, by utilizing beamforming, mmWave allows transmission ranges up to 130 m with data rates up to 385 Mbps and 2 Gbps for a range of 79 m. Moreover, the directional signal can track both moving and stationary objects over a 100 m distance. As a result, beamforming utilized by mmWave can be useful for an IoT application such as smart transportation.

In many IoT applications, an access point (AP) or gateway has to support many devices that periodically send short data packets over a long distance. In fact, LPW and WhiteFi are designed to support such IoT scenarios. Moreover, increased sleep time capability allows LPW-equipped device to sleep for a very long period of time, wake up infrequently to send or receive the data packet, and go back to sleep. Both WiFi and WhiteFi have favorable propagation characteristics, reduced path loss, and better wall-penetration characteristics which make key enabling technologies for many of IoT applications.

DSRC is designed particularly for vehicular communications which supports a large number of vehicles often connected to an AP. The US Department of Transportation recommends using IEEE 802.11p-based DSRC

Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges, Table 2
Summary of candidate wireless technologies for IoT applications

Technology	BLE	LPW	mmWave	DSRC	WhiteFi	LoRa	LTE-M	NB-IoT
Standardization body	Bluetooth SIG	IEEE	IEEE	IEEE	IEEE	LoRa Alliance	3GPP	3GPP
Frequency band	2.4 GHz	Sub-1 GHz	57–64 GHz	5.86–5.92 GHz	TVWS	Sub-1 GHz	7–900 MHz (licensed)	7–900 MHz (licensed)
Channel bandwidth	2 MHz	1–16 MHz	2.16 GHz	10 MHz	6 MHz	~125 kHz	1.4 MHz	200 kHz
Channel access	TDMA	RAW	TDMA or CSMA/CA	CSMA/CA	CSMA/CA	Unslotted ALOHA	SC-FDMA	SC-FDMA
Modulation	GFSK	BPSK to 256 QAM	SC, OFDM	BPSK to 64 QAM	BPSK to 256 QAM	CSS	BPSK to 64 QAM	$\pi/2$ -BPSK, $\pi/4$ -QPSK
Data rate	~1 Mbps	~8 Mbps	~6.76 Gbps	~27 Mbps	~400 Mbps	~38.4 kbps	~1 Mbps	~250 kbps
Coverage	~100 m	~1000 m	~10 m	~1000 m	~1000 m	~5 km	~11 km	~15 km
M2M	Point to point	WiFi Direct	Ad hoc	Ad hoc	Ad hoc	Ad hoc	D2D	D2D

GFSK Gaussian frequency shift modulation, *RAW* restricted access window, *M2M* machine to machine

on light-duty vehicles (http://www.its.dot.gov/factsheets/dsrc_factsheet.htm). But the DSRC technology has limited bandwidth and it can only support a maximum data rate of up to 27 Mb/s.

LoRa, LTE-M, and NB-IoT are three potential long-distance transmission technologies. We note that unlicensed LoRa offers benefits such as power consumption, coverage area, and cost, whereas licensed LTE-M and NB-IoT have advantages in terms of QoS, latency, security, and reliability. LoRa can manage interference and multipath fading well because it uses CSS modulation. LoRa is an asynchronous protocol, whereas LTE-M and NB-IoT are time slotted synchronous protocols. Consequently, LoRa cannot provide the same QoS as LTE-M and NB-IoT do. But this QoS benefit is at the expense of increased cost because the licensed band spectrum is generally a more expensive resource. Hence, the applications that require QoS should choose LTE-M or NB-IoT, whereas applications that do not require high QoS may choose LoRa.

The operating bandwidth of NB-IoT (200 kHz) is narrower than LTE-M (1.08 MHz). The main reason for choosing a narrow band is because LTE-M utilizes the licensed spectrum which is already highly used for human-type traffic and to support a large number of devices.

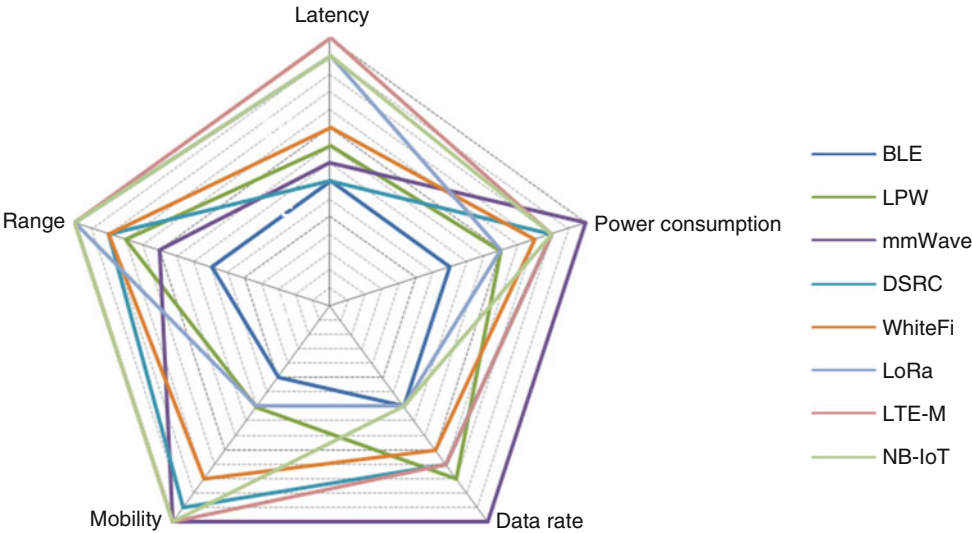
However, a single NB-IoT carrier can support more than 100,000 low data rate devices, and it can increase up to millions of connections if several carriers are used where the LTE-M fails to support.

Figure 2 summarizes the wireless technologies based on the different IoT communication requirements. Table 3 shows a summary of the benefits, limitations of wireless technologies, and their suitability for possible IoT applications.

Case Studies

In this section, we present several case studies that have been reported in the recent literature. We discuss these various previous case studies to present the types of wireless technologies that are being used by different IoT applications.

In Lin et al. (2015), the authors demonstrate an experiment on BLE in smart vehicle IoT application. In this experiment, the authors used Texas Instruments' BLE system-on-chip solution CC2540 (<http://www.ti.com/product/CC2540>) which contains a BLE node and a USB dongle. The BLE node is powered by a CR2032 230 mAh lithium battery. The application layer and a universal asynchronous receiver/transmitter (UART) interface are implemented on an



Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges, Fig. 2
Comparison of potential wireless technologies in terms of IoT communication requirements

Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges, Table 3
Potential wireless technologies for IoT applications

Technologies	Benefits	Limitations	Possible IoT applications
BLE	Point to point, broadcast and mesh support; power saving mechanism is available; large market; widespread adoption in smartphones and consumer electronic devices; support up to 65,000 cell nodes	Crowded spectrum; low data rate; short range; limited bandwidth	Smart Home, Smart Meter, Smart Healthcare, Smart Vehicles
LPW	Low energy consumption by utilizing power saving mechanisms; backhaul connections are supported, adopting a novel hierarchical method that supports up to 8191 devices associated with an AP	Crowded spectrum; attenuation problem	Smart Building, Smart Grid, Smart Industry, Smart Security
mmWave	Worldwide availability, high data rate; high channel bandwidth; mobility support up to 350 km/h	Short range, does not penetrate concrete walls	Smart Vehicles, Smart Transportation, Smart Security, Smart Industry
DSRC	Dedicated but unlicensed only for vehicular communications; mobility support up to 190 km/h	Moderate data rate	Smart Transportation, Smart Vehicles
WhiteFi	Good propagation characteristics; backhaul connections supported; long range	Vulnerable to wireless security attacks	Smart Grid, Smart City
LoRa	Good propagation characteristics; the number of supported devices is around 10^4 per gateway; already deployed and tested; broad coverage area	Lack of expertise to design network	Smart City
LTE-M and NB-IoT	Good propagation characteristics; backhaul connections are unsupported; the number of LTE-M and NB-IoT devices that can be support is around one million and 200,000, respectively; QoS guaranteed; mobility support up to 350 km/h	High cost due to infrastructure and licensed band; depends on current LTE infrastructure; competing against unlicensed LoRa	Smart City, Smart Transportation

electronic control unit (ECU) inside the vehicle. The USB dongle is connected to the ECU through the UART and its adaptation layer serves as the interface between the host and the ECU. Finally, the average current consumptions during connection event and sleep state are obtained 10.655 and 0.9 mA, respectively. Moreover, if the message is transmitted every 2 s, the average current consumption computed from a calculation is 0.013 mA. Finally, the authors estimated that the 230 mAh coin battery life could last around 2 years.

The authors in Adame et al. (2014) present the key features of the IEEE 802.11ah LPW amendment which relates to the decreasing of energy consumption in the MAC layer. They conducted a performance evaluation using MATLAB. They used a one-hop fully connected network and the IoT devices were uniformly distributed within the AP coverage for IoT application scenarios such as agriculture, animal monitoring, smart metering, and industrial automation. The aforementioned scenarios used 3500, 250, 15, and 500 devices, respectively; each device sends data every 120, 60, 50, and 180 s, respectively. In addition, for these scenarios, the energy consumptions obtained were 112, 99, 18, and 25 nJ/bit, whereas, the battery lifetimes were estimated to be 2, 4, 6, and 5 years, respectively. Here to calculate the battery lifetime, the authors used a 2200 mAh lithium battery.

Linghe Kong et al. (2017) explore the capability of mmWave communications for intelligent transportation applications by proposing a novel cloud integrated design of a mmWave system. Moreover, they developed a prototype in a 100 m × 15 m area with three road segments on 30–100 vehicles. Here, all road-side infrastructures and vehicles are equipped with mmWave communications to utilize both vehicle to vehicle (V2V) and vehicle to Internet (V2I) communication support. V2V supports the real-time multimedia data sharing among vehicles, whereas V2I utilizes the roadside infrastructure and cloud computing in order to respond to recognized objects and signs. Finally, they observed that the mmWave demonstrates better channel utilization than DSRC. Figure 3 shows the performance

results of their prototype. In this case, as the number of wireless links used by vehicles increases, both DSRC and mmWave yield a steady linear increase only up to a certain point. This is mainly due to radios in each other's interference coverage area cannot build new wireless links.

Centenaro et al. (2016) described an experiment of IoT networks (smart city) based on LoRa. This experiment was conducted in the Italian city of Padova which has a lot of buildings and shopping malls. In such a harsh propagation condition, the experimental test confirmed that LoRa can support satisfactory long range coverage up to 2 km. Finally, the authors estimated that 30 LoRa gateways would be enough to cover the entire city of Padova which is about 100 square km. Another LoRa empirical test-bed was built (Augustin et al. 2016) and a performance evaluation was carried out in a suburban area of Paris with low-rise residential buildings. About 10,000 data packets were transmitted from end devices to the LoRa gateway and the signal strength of the received packets was measured. From this field trial, the results obtained showed that LoRa was able to cover a gateway around 3 km radius in a suburban area.

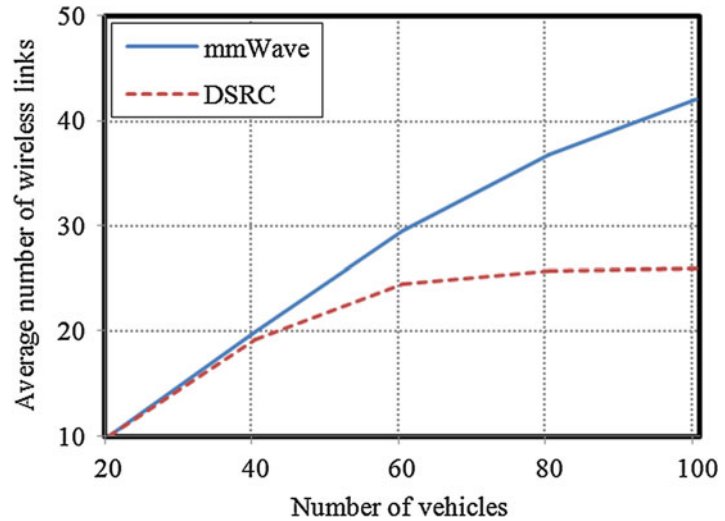
Open Issues

This section presents some open challenges and identifies future research opportunities.

In various IoT scenarios, applications need different QoS requirements in terms of data rate, latency, and reliability. To fulfill QoS requirements becomes challenging particularly in unlicensed bands when there are a large number of connections. Hence, future research needs to focus on the development of QoS-aware protocols that can efficiently support QoS requirements while increasing the coverage area and the number of connections. As we have discussed previously, many IoT applications are delay-sensitive. In the future, new techniques such as D2D communications and local cloud or edge computing have to be carefully integrated to reduce end-to-end communication delays. Moreover, some applications (for example smart

Emerging Wireless Technologies for Internet of Things Applications: Opportunities and Challenges, Fig. 3

The result of the number of vehicles vs the average number effective wireless links (Kong et al. 2017)



E

health) are not only latency sensitive but also data packet loss sensitive. On one hand, in these applications, lost or corrupt data packets for the unreliable network will have serious consequences. On the other hand, some IoT devices are too resource limited to retransmit unacknowledged data. Hence, the reliability issues will have to be defined also.

Future IoT wireless networks will face various security threats such as jamming attack, eavesdropping attack, malicious gateway, or AP during communications. Hence, wireless security techniques need to be improved by considering large-scale connectivity in the future IoT ecosystem. Moreover, security solutions incur additional energy consumption which is a concern in IoT scenarios given that many devices are resource-constrained. Consequently, ensuring security while minimizing the energy consumption needs to be further investigated. We need novel security solutions that can adequately balance the trade-off between security and energy consumption.

We have already mentioned earlier that many IoT devices are battery-powered with very limited processing capability and some of the devices will have to depend on energy harvesting sources. However, to enable large-scale IoT deployments, very low power technologies are required. Although some such technologies have been developed, they are still at an early stage

and still need to be fully tested, evaluated to analyze their suitability. Moreover, emerging low power listening techniques need to be explored which can minimize the power consumption by idle listening.

The amount of generated data in IoT will continue to grow further stressing our communication infrastructures. To manage this exponential growth of IoT data efficiently, we need to develop new networking approaches that can handle a high volume of network traffic efficiently.

Finally, as the licensed spectrum is not free and users need to pay for using it, new policies, economic and pricing models compatible with emerging IoT markets need to be developed when deploying licensed technologies into IoT applications.

Conclusion

The deployment of IoT applications is accelerating with the help of wireless technologies. In this work, we have presented the most promising and viable wireless solutions suitable for IoT applications. We have analyzed the communication requirements of IoT and identified appropriate wireless technologies that can meet those requirements. Several open issues remain to be addressed which we have discussed. We hope this work will be helpful in guiding design-

ers and researchers on the wireless technology choices available for massive IoT deployments. A discussion of power saving mechanisms, communication security, and more deployment case studies will be undertaken as part of our future works.

References

- Adame T, Bel A, Bellalta B, Barcelo J, Oliver M (2014) IEEE 802.11 ah: the WiFi approach for M2M communications. *IEEE Wirel Commun* 21:144–152
- Augustin A, Yi J, Clausen T, Townsley WM (2016) A study of LoRa: long range & low power networks for the internet of things. *Sensors* 16:1466
- Bazzi A, Masini BM, Zanella A, Thibault I (2017) On the performance of IEEE 802.11 p and LTE-V2V for the cooperative awareness of connected vehicles. *IEEE Trans Veh Technol* 66:10419–10432
- Centenaro M, Vangelista L, Zanella A, Zorzi M (2016) Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios. *IEEE Wirel Commun* 23:60–67
- Cisco (2018) Cisco report. https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf. Accessed Oct 2018
- Harris AF III, Khanna V, Tuncay G, Want R, Kravets R (2016) Bluetooth low energy in dense IoT environments. *IEEE Commun Mag* 54:30–36
- Hassan NU, Tushar W, Yuen C, Kerk SG, Oh SW (2017) Guaranteeing QoS using unlicensed TV White Spaces for smart grid applications. *IEEE Wirel Commun* 24:18–25
- Hoymann C, Astely D, Stattin M, Wikstrom G, Cheng J-F, Hoglund A et al (2016) LTE release 14 outlook. *IEEE Commun Mag* 54:44–49
- Ligo AK, Peha JM, Ferreira P, Barros J (2018) Throughput and Economics of DSRC-Based Internet of Vehicles. *IEEE Access* 6:7276–7290
- Kong L, Khan MK, Wu F, Chen G, Zeng P (2017) Millimeter-wave wireless communications for IoT-cloud supported autonomous vehicles: overview, design, and challenges. *IEEE Commun Mag* 55:62–68
- Lauridsen M, Gimenez LC, Rodriguez I, Sorensen TB, Mogensen P (2017) From LTE to 5G for connected mobility. *IEEE Communications Magazine* 55:156–162
- Lin J-R, Talty T, Tonguz OK (2015) On the potential of bluetooth low energy technology for vehicular applications. *IEEE Commun Mag* 53:267–275
- Miao Y, Li W, Tian D, Hossain MS, Alhamid MF (2018) Narrowband Internet of Things: Simulation and Modeling. *IEEE Internet of Things Journal* 5:2304–2314
- Navarro-Ortiz J, Sendra S, Ameigeiras P, Lopez-Soler JM (2018) Integration of LoRaWAN and 4G/5G for the Industrial Internet of Things. *IEEE Communications Magazine* 56:60–67
- Park M (2015) IEEE 802.11 ah: sub-1-GHz license-exempt operation for the internet of things. *IEEE Commun Mag* 53:145–151
- Qiu T, Chen N, Li K, Atiquzzaman M, Zhao W (2018) How can heterogeneous internet of things build our future: A survey. In: *IEEE Communications Surveys & Tutorials*, vol. 20, pp 2011–2027
- Raza U, Kulkarni P, Sooriyabandara M (2017) Low power wide area networks: an overview. *IEEE Commun Surv Tutorials* 19:855–873
- Xu J, Yao J, Wang L, Ming Z, Wu K, Chen L (2018) Narrowband internet of things: Evolutions, technologies, and open issues. *IEEE Internet of Things Journal* 5:1449–1462
- Yilmaz T, Gokkoca G, Akan OB (2016) Millimetre wave communication for 5G IoT applications. In: *Internet of things (IoT) in 5G mobile technologies*. Springer, Cham, pp 37–53

Encrypted Search

► [Encrypted Search for Medical Data](#)

Encrypted Search for Medical Data

Xingliang Yuan

Faculty of Information Technology, Monash University, Melbourne, VIC, Australia

Synonyms

[Encrypted search](#); [Privacy-preserving healthcare service](#)

Definition

Encrypted search over medical data is a set of versatile techniques that allow the data storage server to directly search over different types of encrypted medical data without decryption. Only the data owner and authorized clients with the decryption key can access the data and search results in cleartext.

Historical Background

In the modern age, medical data is digitalized to improve quality of care, promote evidence-based medicine, and facilitate trace of diagnosis and treatment. Despite the benefits expected from it, the data privacy and security issue is still one of the greatest obstacles that hinders the development of digital healthcare systems and services. As surveyed by Fernández-Alemán et al. (2013), over millions of personal medical data records are leaked each year. Because medical data contains personal and private information of patients, the disclosure of such sensitive data raises serious concerns across the entire society. To mitigate the issue, laws and regulations are formulated. For example, the recent HITECH US Health Care Law requires the responsible party to report breaches involving more than 500 patients (Blumenthal 2010). Unfortunately, those policies hardly prevent from massive data breach incidents nowadays. Unauthorized disclosure of medical data can be caused by both unintentional behavior (e.g., software vulnerability or misconfiguration) and intentional behavior (e.g., trading personal information for profits). Also, the threats can be originated from both external and internal entities.

Conventional approaches of protecting medical data are classified into three different techniques, access control, anonymization, and transparent encryption. For a given data item, access control policies specify the privilege scope of the data access at the user or group levels. However, this approach cannot guarantee data confidentiality. No matter how strict access control policies are, data is stored in cleartext. The administrator of the data storage can view the data without any limits. Besides, once a hacker compromises the data storage, all data records are exposed. Another approach is anonymization; patients' identity information is extracted from their medical records before publishing. However, the linkability between patient identities and sanitized data records is possible to be recovered through some disambiguation process if some background or side information is exploitable. To guarantee

strong protection for medical data, encryption is a common practice. Off-the-shelf database systems offer transparent encryption; data is encrypted by the data storage server once the data is written to the database. Unfortunately, this approach relies on the server (not the data owners) to encrypt data for client transparency. Therefore, threats of stealing data still exist as long as the hacker could control the server and obtain the encryption key.

One candidate approach of ensuring confidentiality of medical data is to adopt client-side encryption. Medical data records are encrypted at the client before storing to the database at the server. Without the client's private encryption key, the server always sees the ciphertexts, and the attacker who compromises the server has no access to the data in cleartext throughout its life cycle. But the data storage server now loses the capability of searching and processing medical data and can no longer provide useful healthcare services. How to preserve the utilization of medical data in the encrypted domain becomes an emerging question. As search is a ubiquitous function of data management and analytics, enabling the server to search over encrypted medical data without decryption is demanding for secure and efficient digital health applications. Although there exist generic cryptographic primitives (Shan et al. 2018) that support arbitrary computation such as fully homomorphic encryption and secure multi-party computation, those primitives are too expensive to handle large volumes of encrypted medical data in the real world. Here, we focus on symmetric key-based primitives and introduce three practical applications for searching different types of medical data in different contexts.

Key Applications

Searching Encrypted Textual Medical Records

In digital healthcare, paper-based health records are transformed into electronic health records (EHRs) and stored in the database system for easy management. As mentioned, client-side

encryption is desired for strong protection of data confidentiality. But how to preserve the query functionality of the database over encrypted medical data records becomes an obstacle. After medical data records are inserted to the database, doctors may later search patients' data for diagnosis, prediction, medicine recommendation, and decision-making. Common query functions include quality checks, range queries, joins, and aggregation. Apart from functionality, it is also desired that the adopted techniques should be integrated into the off-the-shelf database management system in a minimally intrusive manner.

To realize the above goals, one may adopt CryptDB, an encrypted relational database management system (RDBMS) designed by Popa et al. (2011). This system implements a set of practical cryptographic primitives to support different SQL query functions, respectively. In particular, CryptDB uses deterministic encryption (DET) for equality check, because DET ciphertexts preserve the quality of the underlying plaintexts. It also uses order-preserving encryption (OPE) for range queries, because OPE ciphertexts preserve the order relations of the underlying numeric data values. On the other hand, proxy re-encryption is used to support the above queries over multiple data attributes encrypted via different keys. The server will transform the ciphertext encrypted from one key to another before equality check and range match. Moreover, additive homomorphic encryption is adopted to support aggregation on encrypted data values.

To improve security and efficiency, CryptDB implements a technique called onion encryption. The values in each data record are encrypted in a layered manner. Based on the related query functions on a data value, it is first encrypted via property-preserving encryption schemes like DET or OPE and then encrypted via randomized encryption. As a result, the server can save some space cost of storing the same data value yet encrypted via different encryption schemes. In addition, if the hacker could only obtain the snapshot of the database, all data values are still strongly protected with semantic security. No

partial information such as equality and order relations is revealed.

Searching Encrypted Medical Images

Digital medical data is not only in the textual format but can also be depicted as images such as magnetic resonance imaging (MRI). In this application context, exact-match queries rarely exist, while similarity queries are more desired. Finding similar medical images in some database with respect to a given query image of a patient is crucial for doctors to conduct disease diagnosis and further decision of treatment. In practice, images are represented as high-dimensional vectors (features), and the similarity of images can normally be quantified by a certain distance between those high-dimensional vectors. To support similarity search over encrypted medical images, one naive approach is to employ secure protocols supporting distance computation on ciphertexts. However, computing pairwise distances between the query image and each image in the database is not scalable even in the plaintext system.

To solve this problem in a secure and efficient manner, Yuan et al. (2015) propose a protocol that systematically integrates the advances in similarity search and encrypted search. Their idea starts from locality-sensitive hashing (LSH) (Andoni and Indyk 2008), an approximate and randomized algorithm for fast similarity search. LSH functions hash high-dimensional vectors in a way that the close ones collide with much higher probability than the distant ones. As a result, the problem of similarity search is transformed to exact matching of LSH hash values derived from image vectors. To enable such matching in the encrypted domain, they further adopt searchable symmetric encryption (SSE) (Curtmola et al. 2011), a practical primitive supporting keyword search in sublinear time complexity. Their methodology is to treat LSH hash values as keywords.

The above direct combination is not practically deployable because of the performance bottleneck raised by the imbalance of LSH. Each

vector needs to be hashed by several LSH functions for accuracy assurance, and one or several query LSH hash values might match a large number of candidate image vectors, leading to high I/O and long latency. To address this issue, they devise a tailored encrypted index from the scratch. This design leverages high-performance hash-based structures, including multiple-choice hashing, cuckoo hashing, and open addressing, and incorporates them into the security framework of SSE. Specifically, they apply pseudo-random functions to protect sensitive LSH hash values, use symmetric encryption to encrypt the index content, and implement the above hash-based optimizations in a random fashion. The resulting encrypted index achieves a high load factor, constant search time, controllable trade-off of accuracy, and privacy preservation for medical image vectors.

Searching Encrypted Streaming Medical Data Signals

Another common type of medical data is generated from medical devices and sensors, because healthcare applications demand routine monitoring to capture a deluge of high-quality information of patients' pulses and motions. For example, a transceiver continuously monitors the physiological signal of a patient, and its sensor can produce up to 31 GB data every day (Yuan et al. 2016). To manage a large scale of medical data signals, cloud-based services are currently used which can link millions of sensors and process data acquired from them. Again, exposing them in a cloud environment may violate regulations of data privacy and cause unauthorized data access. Accordingly, client-side encryption has to be enforced before medical data signals are inserted to the cloud server. Note that those signals are normally generated from sensors in a streaming fashion. Then how to search and process those encrypted and streaming medical data signals becomes even more challenging.

To tackle this challenge, Yuan et al. (2016) propose a privacy-assured framework that can

fast index encrypted medical data signals generated at a high rate. Their first key observation is that unified data sampling and compression techniques like compressive sensing (Candès and Wakin 2008) are being increasingly integrated into hardware sensors to accelerate the acquisition of medical data and reduce the computation and energy consumption during data acquisition. In the meanwhile, one salient feature of compressive sensing is that the Euclidean norm of raw data signals is still preserved in the measurement space of compressed samples. Therefore, doctors are still able to utilize and search over those compressed medical data samples for high-value results, e.g., similar samples. Under this context, the proposed framework introduces a gateway server operated by the trustworthy data owner. It continuously collects data samples from sensing devices via some certain compressive sensing algorithm. But how to encrypt and fast index those samples before sending to the cloud for later use is nontrivial.

This framework resorts to a high-performance encrypted index for similarity search proposed in Yuan et al. (2015). To achieve a high indexing throughput for streaming medical data signals, a new algorithm is designed to construct the above encrypted index with multi-threading support, i.e., enabling parallel insertion while guaranteeing thread safety. In particular, an efficient locking technique called optimistic locking is integrated to the security framework of SSE and the advanced LSH indexing techniques. The proposed concurrent indexing algorithm achieves the optimal number of locks for write operations and avoids intensive use of locks for read operations when seeking multiple index buckets in each insertion. Based on this, this framework includes a secure bulk update mechanism to index continuously generated medical data samples in a batch manner.

Future Directions

To better support privacy-preserving query services for medical data, there are several

challenges yet to be addressed. From the perspective of service scalability, existing approaches and systems focus on the centralized setting where one single server handles and searches all encrypted medical data records. But in practice, a massive amount of medical data is distributed in a cluster of servers. How to enable distributed computing in this domain is demanding. One potential approach is to resort to the distributed framework of the distributed data store and enable parallel query processing in the encrypted domain (Liu et al. 2017). From the perspective of enriching functionality, supporting more advanced computation tasks over encrypted medical data can improve service of quality in digital healthcare while preserving the patient privacy. Recent studies investigate medical image reconstruction (Wang et al. 2013; Shan et al. 2018) and similarity join queries (Yuan et al. 2017), while complicated data analytic functions such as outlier detection in time-series medical data signals and classification over encrypted medical images also need to be explored. From the perspective of security, recent attacks against encrypted search (Naveed et al. 2015; Cash et al. 2015) demonstrate that real-world attackers can still compromise the privacy of the encrypted database if they have access to some background knowledge of the original dataset. The side information captured from the query process can be exploitable with the background knowledge to efficiently recover the plaintext dataset and queries. How to address those new threats efficiently with theoretical guarantee is a challenging and emerging problem in the near future. Otherwise, questions might be raised on whether those primitives and systems supporting encrypted search can securely be adopted for sensitive medical data.

References

- Andoni A, Indyk P (2008) Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun ACM* 51(1):117–122
- Blumenthal D (2010) Launching HITECH. *N Engl J Med* 362(5):382–385
- Candès EJ, Wakin MB (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25(2):21–30
- Cash D, Grubbs P, Perry J, Ristenpart T (2015) Leakage-abuse attacks against searchable encryption. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (CCS)*, pp 668–679
- Curtmola R, Garay JA, Kamara S, Ostrovsky R (2011) Searchable symmetric encryption: improved definitions and efficient constructions. *J Comput Secur* 19(5):895–934
- Fernández-Alemán JL, Señor IC, Lozoya PÁO, Toval A (2013) Security and privacy in electronic health records: a systematic literature review. *J Biomed Inform* 46(3):541–562
- Liu X, Yuan X, Wang C (2017) Encsim: an encrypted similarity search service for distributed high-dimensional datasets. In: *Proceedings of the 25th IEEE/ACM international symposium on quality of service (IWQoS)*, pp 1–10
- Naveed M, Kamara S, Wright CV (2015) Inference attacks on property-preserving encrypted databases. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (CCS)*, pp 644–655
- Popa RA, Redfield CMS, Zeldovich N, Balakrishnan H (2011) Cryptdb: protecting confidentiality with encrypted query processing. In: *Proceedings of the 23rd ACM symposium on operating systems principles (SOSP)*, pp 85–100
- Shan Z, Ren K, Blanton M, Wang C (2018) Practical secure computation outsourcing: a survey. *ACM Comput Surv (CSUR)* 51(2):31
- Wang C, Zhang B, Ren K, Roveda J (2013) Privacy-assured outsourcing of image reconstruction service in cloud. *IEEE Trans Emerg Top Comput* 1(1):166–177
- Yuan X, Cui H, Wang X, Wang C (2015) Enabling privacy-assured similarity retrieval over millions of encrypted records. In: *Proceedings of the 2015 European symposium on research in computer security (ESORICS)*, pp 40–60
- Yuan X, Wang X, Wang C, Weng J, Ren K (2016) Enabling secure and fast indexing for privacy-assured healthcare monitoring via compressive sensing. *IEEE Trans Multimed* 18(10):2002–2014
- Yuan X, Wang X, Wang C, Yu C, Nutanong S (2017) Privacy-preserving similarity joins over encrypted data. *IEEE Trans Inf Forensics Secur* 12(11):2763–2775

Energy Consumption

- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)

Energy Efficiency of Cellular Networks

Xiaohu Ge¹ and Bin Yang²

¹School of Electronic Information and Communications, Director of China International Joint Research Center of Green Communications and Networking, Huazhong University of Science and Technology, Wuhan, China

²Huazhong University of Science and Technology, Wuhan, China

Synonyms

Energy saving of cellular networks; Green networking; Green radio; Green wireless communications

Definition

Energy efficiency of cellular networks aims to improve energy efficiency and resource efficiency of the cellular networks without compromising the quality of services for the users.

Historical Background

During the last decade, there has been tremendous growth in cellular networks' market. Such unprecedented growth in cellular industry has pushed the limits of energy consumption in wireless networks (Hasan et al. 2011). The rising energy costs and carbon footprint of operating cellular networks have led to an emerging trend of addressing energy efficiency among the network operators and regulatory bodies such as the 3rd Generation Partnership Project (3GPP) (3GPP 2010) and the International Telecommunication Union (ITU) (ITU 2010). In this context, the European Commission has recently started new projects within its Seventh Framework Programme to address the energy efficiency of mobile communication systems, i.e., "Energy

Aware Radio and NeTwork TeCHnologies (EARTH)" (Energy Aware Radio and Network Technologies [Earth](#)), "Towards Real Energy-efficient Network Design (TREND)" (Towards Real Energy-Efficient Network Design [Trend](#)) and "Cognitive Radio and Cooperative strategies for Power saving in multi-standard wireless devices (C2POWER)" (Cognitive Radio and Cooperative Strategies for Power Saving in Multistandard Wireless Devices [c2power](#)). In the future fifth-generation (5G) wireless networks, costs, and energy consumption are expected to, ideally, decrease, but at least they should not increase on a per-link basis (Andrews et al. 2014).

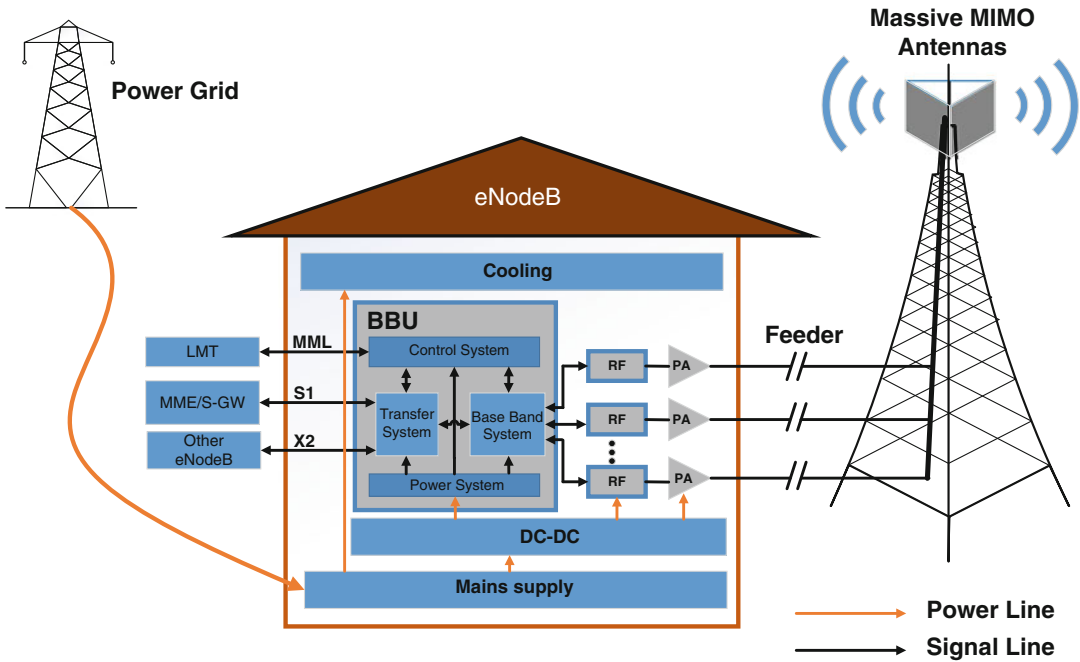
Foundations

Proportions of Energy Consumption

The power consumption of a typical wireless network is mainly for base station (BS), mobile switching, core transmission, data center, retail, etc., in which over 50% is consumed by BSs. In Humar et al. (2011), embodied energy – consumed by all processes associated with the production of equipment – is proposed and is found that embodied energy accounts for a significant proportion of total energy consumption and cannot be neglected. The power consumption by a BS (aka eNodeB in the fourth generation and beyond wireless networks) is illustrated in Fig. 1. Multiple transceivers constitute a BS, and each transceiver consists of a power amplifier, a radio-frequency small-signal transceiver module, a baseband unit (BBU), a DC-DC power supply, an active cooling system, and an AC-DC unit (Ge et al. 2017).

Energy Efficiency Metrics

A classical and widely used metric to evaluate the energy efficiency of telecommunication networks is Bit-per-Joule (Wu et al. 2015). The reciprocal of Bit-per-Joule measurement, namely, energy consumption per delivered information bit, is referred as energy consumption ratio (ECR). A close variant to the Bit-per-Joule metric is Bit-per-Joule-Hz. Instead of system throughput (bits/s), Bit-per-Joule-Hz uses spectral efficiency



Energy Efficiency of Cellular Networks, Fig. 1 The power consumption of a BS (aka eNodeB in the fourth generation and beyond wireless networks)

(bits/s/Hz) as the performance measurement, which is equal to the ratio of spectral efficiency over the total power consumption. In Table 1, a summary of energy efficiency metrics is listed, and several references are recommended.

Methods to Improving Energy Efficiency

Various distinctive approaches can be utilized to reduce energy consumption in a mobile cellular network. Approaches in previous research can be broadly classified into the following five categories.

(1) Improving energy efficiency of hardware components: The energy of cellular networks is mainly consumed by two types of components, i.e., BBU and radio-frequency unit (RFU). Therefore, the way to improve the energy efficiency of hardware components can be divided into two categories. One category is to improve the energy efficiency of BBUs, such as applying latest semiconductor technology to BBUs' chip and designing low overhead algorithm for

BBUs' processing. Another category is to improve the energy efficiency of RFUs, such as applying higher efficiency of the power amplifier to RFUs and changing the connection structure between radio-frequency links and antennas (Zi et al. 2016).

- (2) Turning off components selectively: In practice, the BS density, cell size, and cell capacity are designed to satisfy the requirement of peak traffic, and hence, some BSs are idle at low traffic, which leads to significant energy waste. Selectively turning off components of lightly loaded BSs is proposed to save energy. With aiming at optimization on energy saving only or another energy-related performance trade-offs, several BS switch-off strategies have been proposed from different design perspectives, such as random, distance-aware, load-aware, and auction-based strategies (Jia et al. 2016).
- (3) Optimizing energy efficiency of the radio transmission process: With the rising energy costs and carbon footprint of the cellular

Energy Efficiency of Cellular Networks, Table 1
Energy Efficiency Metrics

Metric	Description	References
Bit-per-Joule	Ratio of effective system capacity over energy consumption	(Ge et al. 2014a, 2015, 2016, 2017)
Bit-per-Joule-Hz	Ratio of effective system capacity over energy consumption in unit bandwidth	(He et al. 2012 Ge et al. 2014b)
Joule-per-Bit	Ratio of energy consumption over effective system capacity	(Badic et al. 2009)
Bit-per-Joule-km ²	Ratio of effective system capacity over energy consumption in a unit area	(Yang et al. 2017)

- networks, optimizing the energy efficiency of the radio transmission process is a critical problem for the future communication systems. Since the power usage of the BS is over 50% of the cellular network power consumption, the optimization of the energy efficiency of the BS is the primary issue to solve. Multiple transceivers constitute a BS, and a transceiver consists of a power amplifier, a radio-frequency small-signal transceiver module, a baseband engine, a DC-DC power supply, an active cooling system, and an AC-DC unit. Furthermore, the power amplifier, the radio frequency, and the baseband engine cost most of the power, so to optimize the energy efficiency of the radio transmission process, the above three mentioned items should be focused (Ge et al. 2017).
- (4) Planning and deploying BSs: Deploying small cells, including micro, pico, and femto cells, in the cellular network. These smaller cells serve small areas with dense traffic with low power-consuming cellular BSs, which are affordable for user deployment and usually support the plug-and-play feature. In contrast to conventional homogeneous macro cell deployment, such heterogeneous deployment reduces energy consumption in the network by shortening the propagation distance between nodes in the network and utilizing higher-frequency bands to support higher data rates (Wu et al. 2015).
- (5) Network resource allocation: Roughly speaking there are two types of resource allocation methods; one is uniform resource

allocation, and the other is nonuniform resource allocation. Uniform resource allocation technologies mainly focus on the fairness between network users to ensure that each user can achieve an acceptable quality of service, while nonuniform resource allocation technologies mainly aim at the enhancement of network performance like sum capacity, throughput, and energy efficiency. Although network resource allocation is quite an old topic, there are still a lot of unsolved problems when fairness and performance are both considered, and issues become even more complicated with energy efficiency constraints.

Key Applications

Information and communication technology (ICT) already represents around 2% of total carbon emissions (of which mobile networks represent about 0.2%), and this is expected to increase every year. In addition to the environmental aspects, energy costs also represent a significant portion of network operators’ overall expenditures (OPEX). While the BSs connected to electrical grid may cost approximately 3000\$ per year to operate, the off-grid BSs in remote areas generally run on diesel power generators and may cost ten times more. To address the challenge of increasing power efficiency in future wireless networks and thereby to maintain profitability, it is crucial to consider various paradigm-shifting technologies,

such as energy efficient wireless architectures and protocols, efficient BS redesign, smart grids, opportunistic network access or cognitive radio, cooperative relaying, and heterogeneous network deployment based on smaller cells.

Cross-References

- [Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices](#)
- [Key Technologies in 4G/LTE Network](#)
- [Robust Games for Power Control in Multi-tier Cellular Networks](#)

Recommended Readings

See References Section.

References

- 3GPP (2010) Tr 32.826, telecommunication management; study on energy savings management (esm), (release 10). Available: <http://www.3gpp.org/ftp/Specs/html-info/32826.htm>
- Andrews JG, Buzzi S, Wan C, Hanly SV, Lozano A, Soong ACK, Zhang JC (2014) What will 5G be? IEEE J Sel Areas Commun 32(6):1065–1082
- Badic B, Farrell TO, Loskot P, He J (2009) Energy efficient radio access architectures for green radio: large versus small cell size deployment. In: Proceedings of the IEEE VTC fall 2009, Anchorage, pp 1–5
- Cognitive Radio and Cooperative Strategies for Power Saving in Multistandard Wireless Devices (c2power). <http://www.ict-c2power.eu/>
- Energy Aware Radio and Network Technologies (Earth). <http://www.ict-earth.eu/>
- Ge X, Yang B, Ye J, Mao G, Li Q (2014a) Performance analysis of poisson-voronoi tessellated random cellular networks using Markov chains. In: Proceedings of the IEEE globecom 2014, Dec 2014, Austin, pp 4635–4640
- Ge X, Cheng H, Guizani M, Han T (2014b) 5G wireless backhaul networks: challenges and research advances. IEEE Netw 28(6):6–11
- Ge X, Yang B, Ye J, Mao G, Wang CX, Han T (2015) Spatial spectrum and energy efficiency of random cellular networks. IEEE Trans Commun 63(3):1019–1030
- Ge X, Tu S, Mao G, Wang CX, Han T (2016) 5G ultra-dense cellular networks. IEEE Wirel Commun 23(1):72–79
- Ge X, Yang J, Gharavi H, Sun Y (2017) Energy efficiency challenges of 5G small cell networks. IEEE Commun Mag 55(5):184–191
- Hasan Z, Boostanimehr H, Bhargava VK (2011) Green cellular networks: a survey, some research issues and challenges. IEEE Commun Surv Tuts 13(4):524–540. FOURTH QUARTER
- He C, Sheng B, Zhu P, You X (2012) Energy efficiency and spectral efficiency tradeoff in downlink distributed antenna systems. IEEE Wirel Commun Lett 1(3):153–156
- Humar I, Ge X, Xiang L, Jo M, Chen M, Zhang J (2011) Rethinking energy efficiency models of cellular networks with embodied energy. IEEE Netw 25(2):40–49
- ITU (2010) Focus group on future networks (fg fn), fg-fn od-66, draft deliverable on “overview of energy saving of networks”. Available: http://www.itu.int/dms_pub/itu-t/oth/3A/05/T3A050000660001MSWE.doc
- Jia H, Chen J, Ge X, Li Q (2016) Switch-off strategy of base stations in HCPP random cellular networks. In: Proceedings of the IEEE ICC 2016, Kuala Lumpur, pp 1–6
- Towards Real Energy-Efficient Network Design (Trend). <http://www.fp7-trend.eu/>
- Wu J, Zhang Y, Zukerman M, Yung EKN (2015) Energy-efficient base-stations sleep-mode techniques in green cellular networks: a survey. IEEE Commun Surv Tuts 17(2):803–826. SECOND QUARTER
- Yang B, Mao G, Ge X, Ding M, Yang X (2017) On the energy-efficient deployment for ultra-dense heterogeneous networks with NLoS and LoS transmissions. Available at: <https://arxiv.org/abs/1712.04201>
- Zi R, Ge X, Thompson J, Wang CX, Wang H, Han T (2016) Energy efficiency optimization of 5G radio frequency chain systems. IEEE J Sel Areas Commun 34(4):758–771

Energy Harvesting

- [Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks](#)

Energy Harvesting and Consumption for Nanoscale Terahertz Communication

- [Energy Modeling for Nanoscale Terahertz Communication](#)

Energy Harvesting Cognitive Radio Sensor Networks

Deyu Zhang

School of Software at Central South University,
Changsha, China

Definitions

Energy harvesting cognitive radio sensor networks operate over idle licensed spectrum using harvested energy. By harvesting energy from the ambient energy sources, such as solar and wind, sensors can achieve sustainable operation without human intervention. In addition, cognitive radio enables sensors to exploit spectrum holes without causing interference to primary users.

Historical Background

Traditional sensor networks consist of miniaturized and low-end sensors powered by batteries with limited capacity. To guarantee the long-term operation, an operator needs to manually replace the depleted battery, which results in considerable maintenance cost. In addition, sensors are networked through the free-of-charge unlicensed spectrum, i.e., the industrial, scientific, and medical (ISM) spectrum, which is becoming more and more congested due to the flourish of Wi-Fi networks. As the miniaturization of energy harvesters and software defined radios, both the academia and industry come to realize that energy harvesting (EH) and cognitive radio (CR) techniques provide promising solution for the resource constraint issue of traditional sensors. The solar panel is first incorporated into sensors to achieve long-lived operation in the early twenty-first century (Lin et al. 2005). After that, numerous researches of EH sensor networks emerge with topics ranging from power management (Kansal et al. 2007), to modeling of EH process (Ku et al. 2015),

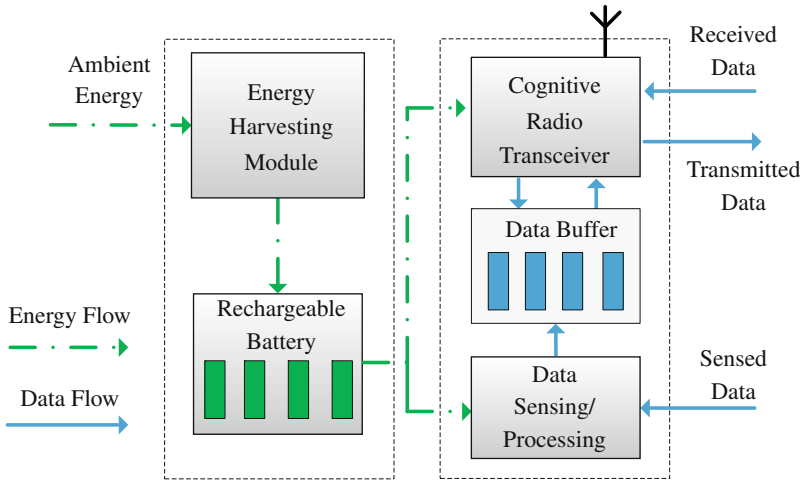
to residual energy-balanced cooperative transmission (Zhang et al. 2016b). To liberate sensors from interferences with other unlicensed networks, Akan et al. (2009) propose cognitive radio sensor networks to enable the opportunistic access of licensed spectrum. Park and Hong (2013) and Park et al. (2013) firstly combine EH and CR techniques in wireless networks. The works investigate the joint schedule of spectrum sensing and access powered by harvested energy. The milestones of EHCRSN appeared in 2016–2017, i.e., Ren et al. (2016); Zhang et al. (2016a, 2017). Ren et al. (2016) optimizes the spectrum access and energy allocation to optimize the network. Without assuming any prior information of the EH process and activities of primary users (PUs), Zhang et al. (2016a) develop an aggregate network utility optimization framework for the design of an online energy and spectrum management based on Lyapunov optimization. Zhang et al. (2017) employ EH-powered spectrum sensors to detect the licensed spectrum to alleviate the spectrum scarcity problem in sensor networks.

Foundations

A typical EHCRSN consists of a sink and numerous sensors with EH and CR modules. The EHCRSN coexists with a primary network which has the privilege to access the licensed spectrum. The sensors use harvested energy to sense data from the AoI and then transmit the sensed data to the sink through the idle licensed spectrum, as shown by green links in Fig. 1. To avoid the interference to PUs, the spectrum detection is mandatory, which can be realized by either a third party system (TPS) or the sensors themselves. Other than sensed data, the sensors may exchange auxiliary information with each other and the sink for spectrum handoff, sensing rate control, routing, etc.

Figure 1 describes a typical EHCR sensor, which consists of five modules: the EH module, rechargeable battery, data sensing/processing module, data buffer, and CR transceiver. The

Deyu Zhang



Energy Harvesting Cognitive Radio Sensor Networks, Fig. 1 Node structure of an EHCR sensor

EH module harvests energy from ambient energy sources and store the energy in the rechargeable battery. The stored energy can be used by the data sensing/processing module which is responsible to collect information from the AoI, and the CR transceiver which detects the licensed spectrum receives and transmits data over the spectrum. The received and sensed data are stored in the data buffer. The ambient energy sources to harvest from determine the type of the EH module. For example, the energy in sunlight can be harvested by solar panel, and the energy in changes of ambient conditions, such as pressure, temperature, and acceleration, can be transformed into electric energy by piezoelectric transducers. To store harvested energy and prevent the impacts of batteries memory effect, the super capacity or lithium-ion battery are considered as competitive candidates for rechargeable battery in the EHCR sensors.

Considering the limitation on the manufacture cost of sensors, the low-cost CR transceivers are desired, which can cover the licensed spectrum of interest, i.e., the spectrum that is accessible to EHCR sensors. Since the price of the CR transceivers tends to increase with the width of the tunable frequency range, it is important to choose the CR transceiver according to the accessed licensed spectrum.

Key Applications

With EH and CR capabilities, EHCRSNs can perform monitoring and surveillance tasks sustainably while overlapping with other unlicensed networks, such as Wi-Fi hotspots and Bluetooth. Therefore, the applicability of EHCRSNs is significantly improved in various real-life scenarios such as smart city, indoor surveillance and localization, and electronic health.

Future Directions

From the evolution of wireless sensor networks, it appears that the future needs will come from the modeling of EH process and PU activities, the joint spectrum sensing and access schemes using harvested energy, as well as the routing in multi-hop EHCRSNs. In the following are some future directions:

- *The Modeling of EH process and PU activities based on real data trace.* This issue remains open mainly due to the lack of comprehensive historical data records in EH process and licensed spectrum usage. In the coming era of Internet of Things, more data records

of the two processes can be complemented by the widely deployed devices with sensing capability, e.g., smartphones. Based on the records, the operator can use the data mining and machine learning techniques to extract the statistical information of EH process and PU activities.

- *Joint Spectrum Detection and Access.* For EHCR sensors, both the exploration and exploitation of licensed spectrum consume energy which makes the coupling of energy and spectrum allocation more complex. The tradeoff exists between the accurate information on PU activity and immediate spectrum access.
- *Resource Allocation in Multi-hop EHCRSNs.* Some sensor networks may consist of thousands of nodes which transmit data to the sink through multi-hop relaying. Considering the signaling overhead of centralized algorithms, distributed algorithms are desired to improve the scalability of EHCRSNs. To utilize the benefits brought by EH and CR capabilities, sensors need to frequently adjust power control, change next-hop relay, and vacate operating channels.

Cross-References

- [Cognitive Radio Networks](#)
- [Wireless Sensor Networks](#)

References

- Akan O, Karli O, Ergul O (2009) Cognitive radio sensor networks. *IEEE Netw* 23(4):34–40
- Kansal A, Hsu J, Zahedi S, Srivastava MB (2007) Power management in energy harvesting sensor networks. *ACM Trans Embed Comput Syst* 6(4). <https://doi.org/10.1145/1274858.1274870>
- Ku ML, Chen Y, Liu KJR (2015) Data-driven stochastic models and policies for energy harvesting sensor communications. *IEEE J Sel Areas Commun* 33(8):1505–1520. <https://doi.org/10.1109/JSAC.2015.2391651>
- Lin K, Yu J, Hsu J, Zahedi S, Lee D, Friedman J, Kansal A, Raghunathan V, Srivastava M (2005) Helimote: enabling long-lived sensor networks through solar energy harvesting. In: *Proceedings of the 3rd international conference on embedded networked sensor systems*, ACM, New York, SenSys '05, pp 309–309, DOI <https://doi.org/10.1145/1098918.1098974>
- Park S, Hong D (2013) Optimal spectrum access for energy harvesting cognitive radio networks. *IEEE Trans Wirel Commun* 12(12): 6166–6179
- Park S, Kim H, Hong D (2013) Cognitive radio networks with energy harvesting. *IEEE Trans Wirel Commun* 12(3):1386–1397. <https://doi.org/10.1109/TWC.2013.012413.121009>
- Ren J, Zhang Y, Deng R, Zhang N, Zhang D, Shen X (2016) Joint channel access and sampling rate control in energy harvesting cognitive radio sensor networks. *IEEE Trans Emerg Top Comput.* <https://doi.org/10.1109/TETC.2016.2555806>
- Zhang D, Chen Z, Awad MK, Zhang N, Zhou H, Shen XS (2016a) Utility-optimal resource management and allocation algorithm for energy harvesting cognitive radio sensor networks. *IEEE J Sel Areas Commun* 34(12):3552–3565
- Zhang D, Chen Z, Zhou H, Chen L, Shen X (2016b) Energy-balanced cooperative transmission based on relay selection and power control in energy harvesting wireless sensor network. *Comput Netw* 104: 189–197
- Zhang D, Chen Z, Ren J, Zhang N, Awad MK, Zhou H, Shen XS (2017) Energy-harvesting-aided spectrum sensing and data transmission in heterogeneous cognitive radio sensor network. *IEEE Trans Veh Technol* 66(1):831–843

Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks

Liang Huang

College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

Synonyms

[Energy harvesting](#); [Scheduling](#); [Wireless sensor networks](#)

Definitions

In wireless sensor networks (WSNs), an *energy harvesting sensor node* harvests ambient energy from the environment and converts it into electrical energy to support continuous WSN oper-

ations without energy depletion. To overcome the dynamic feature of energy harvesting, *energy harvesting sensor node scheduling* jointly regulates data transmission and energy management to achieve reliable and efficient operations in WSNs.

Historical Background

WSNs have been widely deployed in the past decades for various applications, for example, environmental monitoring (Adu-Manu et al. 2018), animal tracking (Szewczyk et al 2004), and smart cities (Bakakeu et al. 2017). They consist of a number of low-cost and battery-operated sensor nodes that transfer sensed data over the air. An important performance metric of WSNs is the lifetime constrained by limited batteries. Over the years, different energy-efficient sensing and communication protocols (Khajuria and Gupta 2015) have been developed to minimize the energy consumption of sensor nodes. However, the sensor node is inevitable to be out of service once its limited battery energy is exhausted. To address this challenge, researchers have explored technologies to power sensor nodes by the ambient energy harvested, such as solar, wind, vibration, thermoelectric power, and radio frequency power, since the early 2000s (Adu-Manu et al. 2018). Due to the dynamic feature of energy harvesting, it is important to carefully design the energy management, especially the communication energy, to guarantee systems performance, such as connectivity, throughput, and delay (Pantazis and Vergados 2007).

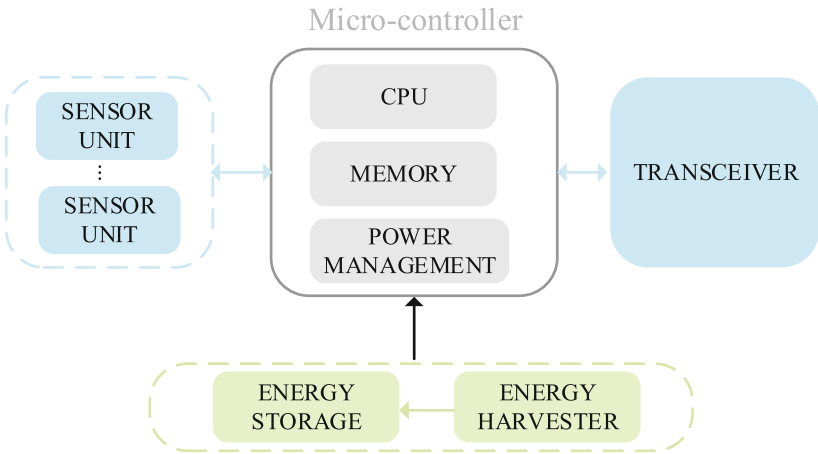
In energy-harvesting WSNs, scheduling policies greatly depends on the harvested energy stored in the sensor node. A mathematical framework for energy harvesting sensor node scheduling is proposed in Lei et al. (2009), where different energy storage states are modeled as discrete-time Markov chains. Data packets with high priorities are preferred to be transmitted when energy storage is low. Under causality constraints on both data arriving process and energy harvesting process, Yang and Ulukus

(2012) studied optimal packet scheduling to minimize the transmission time for data packets by adaptively controlling the transmission rate and power. By considering the resource-limited hardware of energy harvesting sensor nodes, Liang et al. (2018) proposed a threshold-based scheduling policy for energy harvesting sensor networks to maximize its expected reward, which approaches the theoretical offline upper bound.

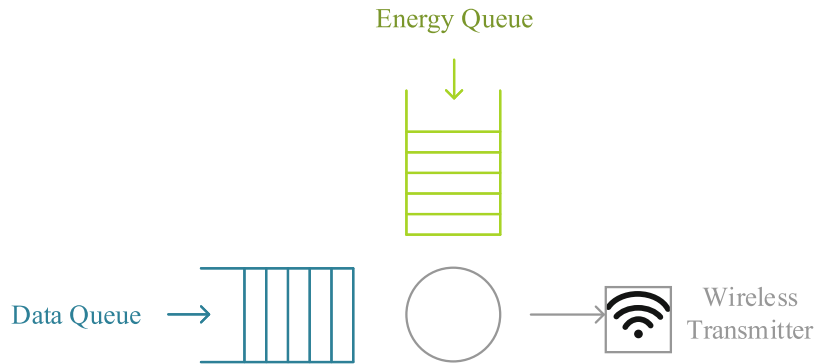
Foundations

An energy harvesting sensor node is composed of a microcontroller, a wireless transceiver, an energy harvester with storage, and multiple sensing units, as illustrated in Fig. 1. The sensor node converts harvested energy into electrical energy and store it in a finite-size energy storage for future data transmissions. Those sensed/generated data packets are stored in a memory structure and will consume some harvested energy during transmission. The hardware of both data memory and energy storage are size-limited. A data packet is transmitted only when there is enough energy stored/harvested. Under this causality constraint, data packets transmission scheduling at energy harvesting sensor node must be carefully managed.

A general system model for energy harvesting sensor node scheduling is presented in Fig. 2, where both data generation-transmission process and energy harvesting-consumption process are modeled as two independent dynamic queueing processes. Each successful transmission of a data packet will gain some rewards, whose value may be evaluated based on different performance metrics of interest (Ku et al. 2016), including transmission completion time, data throughput, outage probability, mean delay, message importance, quality of coverage, etc. An optimal scheduling policy aims to maximize the average transmission reward of the energy harvesting WSNs. Based on whether the knowledge of energy and data arrivals is available or not, packet scheduling policies can be categorized into *offline* scheduling and *online* scheduling:



Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks, Fig. 1 Architecture of an energy harvesting sensor node in wireless sensor networks



Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks, Fig. 2 System model for energy harvesting sensor node scheduling

- The *offline* scheduling assumes full knowledge of energy and data arrivals in the coming future and aims to develop an optimal scheduling policy for energy harvesting WSNs (Yang and Ulukus 2012). The study provides a theoretical optimal scheduling in energy harvesting WSNs by maximizing a certain short-term reward over a finite time horizon. In general, such kind of optimization problems can be solved by convex optimization techniques. The study of *offline* scheduling not only guides the design and deployment of WSNs but also provides an upper bound for further *online* scheduling studies.
- The *online* scheduling only has knowledge of current and past system status and tar-

gets to be implemented in real field deployment of WSNs. In order to perform real-time scheduling, threshold-based scheduling policies are widely adopted for *online* scheduling (Liang et al. 2018). Specifically, for each energy storage state, a corresponding transmission reward threshold is pre-computed and stored in a look-up table. Before each transmission, the sensor node searches for the corresponding threshold based on its energy storage state and only transmits a packet when its reward value is above the threshold. Due to the discrete nature of reward value thresholds, stochastic optimization techniques are used to maximize the long-term rewards of the energy harvesting WSNs.

Key Applications

Energy harvesting sensor node scheduling is fundamental in every energy harvesting WSN, which helps improve system performance metrics of interest, such as throughput, delay, and data importance.

Cross-References

- [Energy Harvesting Technologies in Wireless Sensor Networks](#)
- [Extending WSN Lifetime based on Evolutionary Clustering Algorithm](#)
- [Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes](#)

References

- Adu-Manu KS, Adam N, Tapparello C, Ayatollahi H, Heinzelman W (2018) Energy-harvesting wireless sensor networks (EH-WSNs): a review. *ACM Trans Sens Netw* 14(2):10:1–10:50. <https://doi.org/10.1145/3183338>
- Bakakeu J, Schäfer F, Bauer J, Michl M, Franke J (2017) Building cyber-physical systems – a smart building use case. In: Houbing S (ed) *Smart cities: foundations, principles, and applications*. Wiley, Hoboken, pp 605–639
- Khajuria R, Gupta S (2015) Energy optimization and lifetime enhancement techniques in wireless sensor networks: a survey. In: *International conference on computing, communication automation*, pp 396–402. <https://doi.org/10.1109/CCAA.2015.7148408>
- Ku ML, Li W, Chen Y, Liu KJR (2016) Advances in energy harvesting communications: past, present, and future challenges. *IEEE Commun Surv Tutor* 18(2):1384–1412. <https://doi.org/10.1109/COMST.2015.2497324>
- Lei J, Yates R, Greenstein L (2009) A generic model for optimizing single-hop transmission policy of replenishable sensors. *IEEE Trans Wirel Commun* 8(2):547–551. <https://doi.org/10.1109/TWC.2009.070905>
- Liang H, Li Ping Q, Suzhi B, Zhuoqun X (2018) Adaptive scheduling in energy harvesting sensor networks for green cities. *IEEE Trans Ind Inf* 14(4):1575–1584. <https://doi.org/10.1109/TII.2017.2710313>
- Pantazis NA, Vergados DD (2007) A survey on power control issues in wireless sensor networks. *IEEE Commun Surv Tutor* 9(4):86–107
- Szewczyk R, Polastre J, Mainwaring A, Culler D (2004) Lessons from a sensor network expedition. In: Karl H, Wolisz A, Willig A (eds) *Wireless sensor networks*. Springer, Berlin/Heidelberg, pp 307–322
- Yang J, Ulukus S (2012) Optimal packet scheduling in an energy harvesting communication system. *IEEE Trans Commun* 60(1):220–230

Energy Harvesting Technologies in Wireless Sensor Networks

Yongmin Zhang

University of Victoria, Victoria, BC, Canada

Synonyms

[Ambient power technology](#); [Energy scavenging technology](#); [Power harvesting technology](#); [Wireless recharging technology](#)

Definitions

Energy is an important resource in wireless sensor networks since it determines the operation performance of sensors, i.e., data sensing, data processing, and data transmission. For the long-term applications, the power supply is one of the most important issues should be addressed since a fixed utility supply and/or manual battery changing may not be technically and economically viable. To address the power supply issue with the goal to obtain perpetual and unattended networks, energy harvesting technologies, which can harvest energy from energy sources in the surroundings (i.e., solar power, thermal energy, wind energy, kinetic energy, and electromagnetic energy), have been recently introduced to power the sensor nodes and thus achieve the goal of perpetual network operation. This kind of wireless sensor networks is known as rechargeable sensor networks or energy harvesting sensor networks.

Background

Generally, a wireless sensor network is composed of a large number of static sensor nodes with low-power device and limited processing capabilities to collect environment information. Each sensor node typically has three basic functions: (i) data sensing for acquiring environment information, (ii) data processing for processing the sensed data locally, and (iii) wireless communication for receiving and transmitting data. All of these functions are energy-consuming, especially in the large-scale and long-distance transmission wireless sensor networks. With development of wireless sensor networks, the need to extend the lifetime of wireless sensor networks becomes more and more urgent. Low-energy consumption technologies have been developed to prolong the lifetime of sensor nodes. However, low-energy consumption technologies cannot fundamentally solve the problem of insufficient energy in wireless sensor networks since a lot of applications require the sensor nodes can fulfill their requirements for a long duration (i.e., several months or several years) (Niyato et al. 2007; Shaikh and Zeadally 2016). Furthermore, it is very difficult, even impossible, to replace the batteries due to the difficult and hostile deployment terrain or large amount of sensors. As battery-powered sensor nodes are not feasible for long-term applications, energy harvesting technologies have been recently introduced to achieve the goal of perpetual network operation (Adu-Manu et al. 2018).

Since the energy sources are made up of different forms of ambient energy, there is no one solution that would fit all of applications. To satisfy various requirements of applications, many different types of energy harvesting systems have been designed. Generally, energy harvesting systems consist of: energy collection elements, conversion hardware and power conditioning, and storage devices as shown in Fig. 1. Energy collection elements determine the amount of energy that can be harvested. Conversion hardware and power conditioning convert the harvested energy to the electricity and conditioned to an appropri-

ate form for either charging the system batteries or powering the sensor node directly. Storage devices reserve the unused harvested energy for future use and power the sensor nodes if needed. The power densities of some common energy sources are shown in Table 2. Based on the requirements of applications, how to design the energy harvesting system (includes selections of energy source, energy collecting and conditioning circuit, and energy storage) and its corresponding energy management scheme is important. In the following part, some energy harvesting technologies are introduced based on energy sources and applications.

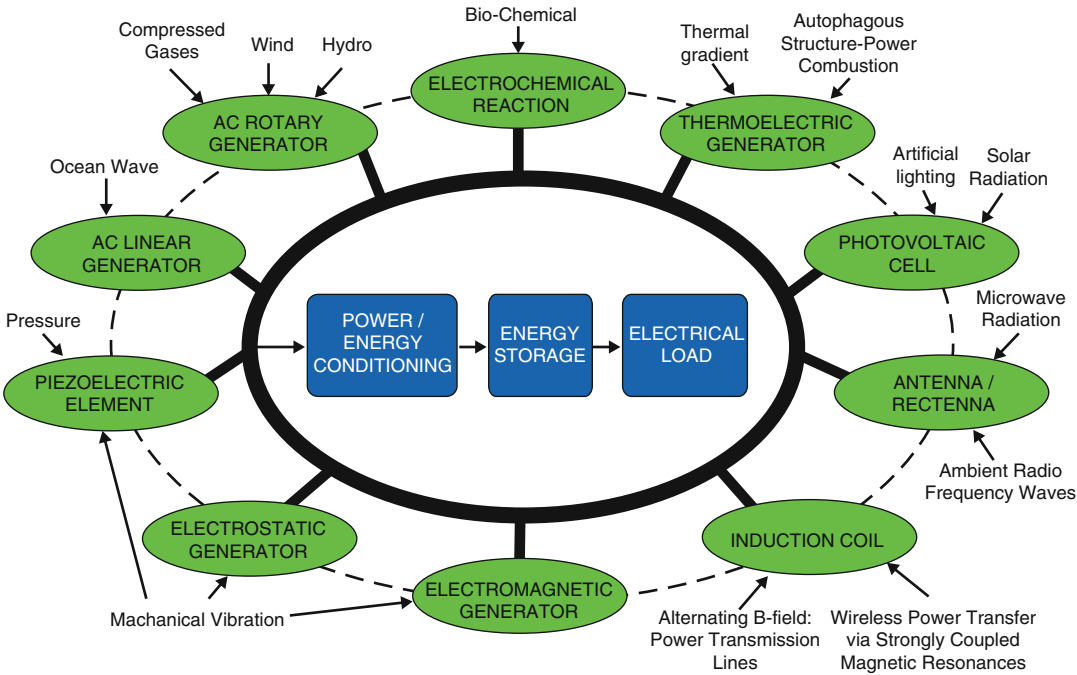
Energy harvesting sensor nodes can harvest ambient energy from sources in the surroundings, and store the harvested energy in the batteries for future use, such that sensor nodes can work perpetually. Normally, different ambient sources have different features and different energy harvesting technologies. For example, solar energy collected by the solar panel is predictable and non-controllable. Radio Frequency (RF) energy collected by the RF receiver is predictable and controllable. Some common energy sources and their corresponding characteristics are shown in Table 1.

Energy Harvesting Wireless Sensor Networks

Some common energy harvesting wireless sensor networks are introduced as follows.

RF-Based Energy Harvesting Wireless Sensor Networks

RF-based energy harvesting technology is one of the most popular energy harvesting technologies due to its predictable and controllable features. For RF-based energy harvesting sensor networks, the sensor nodes collect the energy carried by radio waves and convert it to DC power to power themselves. There are several approaches to convert the RF signals into DC power (i.e., single-stage vs. multistage). The selection of approaches



Energy Harvesting Technologies in Wireless Sensor Networks, Fig. 1 Energy sources and respective transducers (Tan and Panda 2010) © 2010 Tan YK, Panda SK.

Published in (Tan and Panda 2010) under CC BY-NC-SA 3.0 license. Available from: <https://doi.org/10.5772/13062>

Energy Harvesting Technologies in Wireless Sensor Networks, Table 1 Characteristics of various energy sources (Shaikh and Zeadally 2016)

Energy source	Predictable	Unpredictable	Controllable	Non-controllable
RF	*		*	
Solar	*			*
Wind	*			*
Thermal		*	*	
Hydro	*			*
Vibration		*	*	
Pressure		*	*	
StressCstrain		*	*	
Activity		*	*	
Physiological		*		*

* denotes that the energy source has the corresponding feature

depends on the desired application requirements (i.e., power, efficiency, or voltage). There exist several factors that impact the amount of harvested RF energy, including but not limited to the source power, antenna gain, distance between

source and the receiver, and energy conversion efficiency.

In recent years, RF energy harvesting technologies have been explored extensively in an attempt to utilize this energy source to power

Energy Harvesting Technologies in Wireless Sensor Networks, Table 2 The energy densities of some energy sources (Adu-Manu et al. 2018)

Energy sources	Energy-harvesting method	Power density
RF	Electromagnetic conversion	0.1 $\mu W/cm^2$ (GSM)
		0.01 $\mu W/cm^2$ (WiFi)
Solar	Solar cells	$\leq 10 \mu W/cm^2$ (indoors)
		15 mW/cm^2 (outdoors)
Wind and Hydro	Electromechanical conversion	16.2 $\mu W/cm^3$
Thermal (Body heat)	Thermoelectric	40 $\mu W/cm^2$

wireless sensor networks for environmental monitoring applications. A survey on wireless networks using RF energy harvesting has been summarized in Lu et al. (2015) and Alsaba et al. (2018). Furthermore, there exist a lot of works on circuit design, communication protocols, and operation design (Shu et al. 2017; Ren et al. 2018; Mouapi and Hakem 2018). RF-based energy harvesting wireless sensor networks are suitable for the application in urban and suburban areas due to the large amount of ambient energy sources and the controllable feature.

Solar-Based Energy Harvesting Wireless Sensor Networks

Solar-based energy harvesting technology is another popular energy harvesting technology due to its predictable and abundant features in the environment. There are two main approaches for converting sunlight into usable energy: thermal conversion and photovoltaic conversion. Photovoltaic conversion is more popular due to its higher-power density. There are a lot of factors that affected solar energy efficiency, i.e., cloud cover, sun intensity, relative humidity, and heat buildup. Due to the limitations of solar energy harvesting systems during the night, developers must ensure the highest possible efficiency during

day light hours to guarantee the viability of solar power.

Recently, Solar-based wireless sensor networks have been widely explored. A survey on the solar-based energy harvesting applications can be found in Sharma et al. (2018). Also, there exist a lot of works on the design of sensor nodes and the optimization algorithms/protocols to improve the network performance (Zhang et al. 2016, 2017; Wang et al. 2018). The solar-based energy harvesting wireless sensor networks are suitable for environmental monitoring applications due to its abundance in the environment.

Wind-Based Energy Harvesting Wireless Sensor Networks

Wind-based energy harvesting technology is a good alternative power source due to its high-energy density and abundance in the environment. As a form of kinetic energy, it can be collected by (micro) wind turbine to power the sensor nodes. Due to its unpredictable and non-controllable features, the energy supply always is fragile and unreliable. There exists a lot of works on the design of wind-based energy harvesting system and optimization of the network operation (Sardini and Serpelloni 2011; Ma et al. 2014; Porcarelli et al. 2015). Due to the non-controllable and unpredictable features of wind energy, to ensure the network performance, wind energy always is designed with other energy sources; some applications can be found in the following hybrid energy-harvesting systems.

Thermal-Based Energy Harvesting Wireless Sensor Networks

Thermal energy harvesting is based on the fact that when there is a temperature difference between two conducting materials, an electric current is generated. For Thermal-based sensors, both the environmental requirements and the thermoelectric materials are demanding. There exist some environmental studies on thermal powered sensor nodes, i.e., a thermoelectric generator (TEG) was mounted to a sheeps collar

to collect the temperature gradient between the sheep body and the ambient environment (Woiias et al. 2014), thermal energy harvesting between the air/water interface for powering sensor nodes (Zhou et al. 2014). Typically, thermal harvesting efficiency is 5–6%, which is low and becomes a major hindrance for its widespread adoption. Thermal energy harvesters are more suitable for large-scale power generation such as steam turbines (Shaikh and Zeadally 2016).

Human-Based Energy Harvesting Wireless Sensor Networks

Human-based energy harvesting technologies can be categorized as activity-based energy harvesting and inherent physiological parameters-based energy harvesting. Human-based energy harvesting sensor nodes can harvest energy from human activities (i.e., locomotion or changes in finger position, body heat, blood flow, movement, and heartbeat) and convert the collected energy into the available power to maintain their operations. In recent years, Wireless Body Area Network (WBAN) has been receiving a lot of attention recently (Mukhopadhyay 2015; Thielen et al. 2017; Rault et al. 2017).

Hybrid Energy-Harvesting Systems

Due to different features of different energy sources, combining multiple types of energy harvesters in a single sensor node can achieve better performance. There exists several common hybrid energy harvesting systems:

1. Solar/Thermal Systems in organic fertilizer plant, which consists of solar panels and TEG as energy converters (Chottirapong et al. 2015).
2. Solar/Thermal/Electromagnetic Systems to improve the performance of autonomous wireless network nodes (Virili et al. 2015, 2016).
3. Solar/wind Systems to build energy-autonomous sensing systems (Cammarano et al. 2016; Spenza et al. 2012).

References

- Adu-Manu KS, Adam N, Tapparello C, Ayatollahi H, Heinzelman W (2018) Energy-harvesting wireless sensor networks (eh-wsns): a review. *ACM Trans Sens Netw* 14(2):10
- Alsaba Y, Rahim SKA, Leow CY (2018) Beamforming in wireless energy harvesting communications systems: a survey. *IEEE Commun Surv Tutor* 20(2):1329–1360
- Cammarano A, Petrioli C, Spenza D (2016) Online energy harvesting prediction in environmentally powered wireless sensor networks. *IEEE Sensors J* 16(17):6793–6804
- Chottirapong K, Manatrinon S, Dangsakul P, Kwankeow N (2015) Design of energy harvesting thermoelectric generator with wireless sensors in organic fertilizer plant. In: 2015 IEEE 6th international conference of information and communication technology for embedded systems (IEEE IC-ICTES), pp 1–6
- Lu X, Wang P, Niyato D, Kim DI, Han Z (2015) Wireless networks with rf energy harvesting: a contemporary survey. *IEEE Commun Surv Tutor* 17(2):757–789
- Ma L, Li B, Yang ZB, Du J, Wang J (2014) A new combination prediction model for short-term wind farm output power based on meteorological data collected by WSN. *Int J Control Autom* 7:171–180
- Mouapi A, Hakem N (2018) A new approach to design autonomous wireless sensor node based on rf energy harvesting system. *Sensors* 18(1):133
- Mukhopadhyay SC (2015) Wearable sensors for human activity monitoring: a review. *IEEE Sensors J* 15(3):1321–1330
- Niyato D, Hossain E, Rashid MM, Bhargava VK (2007) Wireless sensor networks with energy harvesting technologies: a game-theoretic approach to optimal energy management. *IEEE Wirel Commun* 14(4):90–96
- Porcarelli D, Spenza D, Brunelli D, Cammarano A, Petrioli C, Benini L (2015) Adaptive rectifier driven by power intake predictors for wind energy harvesting sensor networks. *IEEE J Emerging Sel Top Power Electron* 3(2):471–482
- Rault T, Bouabdallah A, Challal Y, Marin F (2017) A survey of energy-efficient context recognition systems using wearable sensors for healthcare applications. *Pervasive Mob Comput* 37:23–44
- Ren J, Hu J, Zhang D, Guo H, Zhang Y, Shen X (2018) Rf energy harvesting and transfer in cognitive radio sensor networks: opportunities and challenges. *IEEE Commun Mag* 56(1):104–110
- Sardini E, Serpelloni M (2011) Self-powered wireless sensor for air temperature and velocity measurements with energy harvesting capability. *IEEE Trans Instrum Meas* 60(5):1838–1844
- Shaikh FK, Zeadally S (2016) Energy harvesting in wireless sensor networks: a comprehensive review. *Renew Sust Energ Rev* 55:1041–1054

- Sharma H, Haque A, Jaffery ZA (2018) Solar energy harvesting wireless sensor network nodes: a survey. *J Renewable Sustainable Energy* 10(2):023,704
- Shu Y, Shin KG, Chen J, Sun Y (2017) Joint energy replenishment and operation scheduling in wireless rechargeable sensor networks. *IEEE Trans Ind Inf* 13(1):125–134
- Spenza D, Petrioli C, Cammarano A (2012) Pro-energy: a novel energy prediction model for solar and wind energy-harvesting wireless sensor networks. In: 2012 9th IEEE international conference on mobile ad-hoc and sensor systems (IEEE MASS 2012), pp 75–83
- Tan YK, Panda SK (2010) Review of energy harvesting technologies for sustainable WSN. In: Tan YK (ed) *Sustainable wireless sensor networks*. InTech
- Thielen M, Sigrist L, Magno M, Hierold C, Benini L (2017) Human body heat for powering wearable devices: from thermal energy to application. *Energy Convers Manag* 131:44–54
- Virili M, Georgiadis A, Collado A, Niotaki K, Mezzanotte P, Roselli L, Alimenti F, Carvalho NB (2015) Performance improvement of rectifiers for wpt exploiting thermal energy harvesting. *Wirel Power Transf* 2(1):22–31
- Virili M, Georgiadis A, Mira F, Collado A, Alimenti F, Mezzanotte P, Roselli L (2016) Eh performance of an hybrid energy harvester for autonomous nodes. In: 2016 IEEE topical conference on wireless sensors and sensor networks (IEEE WiSNet), pp 71–74
- Wang C, Li J, Yang Y, Ye F (2018) Combining solar energy harvesting with wireless charging for hybrid wireless sensor networks. *IEEE Trans Mob Comput* 17(3):560–576
- Woiass P, Schule F, Bäumke E, Mehne P, Kroener M (2014) Thermal energy harvesting from wildlife. *J Phys Conf Ser* 557:012084
- Zhang Y, He S, Chen J (2016) Data gathering optimization by dynamic sensing and routing in rechargeable sensor networks. *IEEE/ACM Trans Networking* 24(3):1632–1646
- Zhang Y, He S, Chen J (2017) Near optimal data gathering in rechargeable sensor networks with a mobile sink. *IEEE Trans Mob Comput* 16(6):1718–1729
- Zhou G, Huang L, Li W, Zhu Z (2014) Harvesting ambient environmental energy for wireless sensor networks: a survey. *J Sens* 2014, pp 1–20

Energy Modeling for Nanonetworks in the Terahertz Band

- [Energy Modeling for Nanoscale Terahertz Communication](#)

Energy Modeling for Nanoscale Terahertz Communication

Xin-Wei Yao

Zhejiang University of Technology, Hangzhou, China

Synonyms

[Energy harvesting and consumption for nanoscale terahertz communication](#); [Energy modeling for nanonetworks in the terahertz band](#)

Definitions

Energy modeling for nanoscale terahertz communication is required to investigate the relationship between energy harvesting and consumption in nanoscale terahertz communication and analyze the trade-off between them by jointly considering the constraints of nanonetworks and the peculiarities of terahertz communications, in order to achieve perpetual nanonetworks.

Historical Background

Nanotechnology is enabling the development of novel devices in the order of a few cubic micrometers in size, which can accomplish only very simple tasks (Akyildiz and Jornet 2010). Due to the very limited abilities of individual nanodevice, communication among nanodevices will expand the potential applications of single nanodevice through collaboration. The resulting nanonetworks will be able to boost the range of applications of nanotechnology in biomedical, environmental, and military fields, such as intrabody health monitoring systems, distributed air pollution control, and nanosensor networks (Krishnaswamy et al. 2010). Two main alternatives for communication in the nanoscale have been considered, i.e., molecular communication and electromagnetic

communication, respectively. In detail, molecular communication is defined as the transmission and reception of information encoded in molecules, while electromagnetic communication is defined as the transmission and reception of electromagnetic radiation from nanodevices using novel nanomaterials (Akyildiz and Jornet 2010). In this chapter, electromagnetic communication among nanodevices in the terahertz band is considered and introduced.

According to the extremely small size of nanodevices, scaling a metallic antenna down to a few hundred nanometers would impose the use of very high operating frequencies, thus, drastically limiting the communication range of nanodevices. Alternatively, graphene and its derivatives, such as carbon nanotubes and graphene nanoribbons, can be used to develop nanoantennas able to radiate at much lower frequencies. This frequency band, named Terahertz (THz) Band, spans the frequency range from 0.1 to 10.0 THz (Jornet and Akyildiz 2011). For the time being, due to the strict constraints of a single nanodevice in terms of size and energy, no integrated technology has been presented to generate a high-power carrier frequency in the THz band. As a result, the traditional wireless communication mechanisms on the basis of continuous transmission signals may be not suitable for nanonetworks with limited hardware. Inspired by the huge bandwidth provided by the THz band, new pulse-based communication mechanisms have been envisioned as the candidates for nanonetworks by the exchange of a few femtoseconds-long pulses. Moreover, the energy requirement on the transmission is relaxed by distributing short pulses over the time rather than continuous signals (Yao et al. 2015).

On the other hand, the very limited energy stored in the nanobatteries is another major challenge in nanonetworks. Therefore, energy harvesting system has been considered to provide energy for nanodevices (Wang 2008). Unfortunately, conventional energy harvesting systems, such as solar energy, or wind power, cannot be adopted owing to the extremely small size of nanodevices, just several cubic micrometers (Sude-

valayam and Kulkarni 2011). Recently, piezoelectric nanogenerators have been proposed to recharge the nanodevices (Yao et al. 2018). In addition, some works considered the energy consumption process of electromagnetic communication in the THz band and even for multi-hop nanonetworks in the aspect of MAC-layer or Network-layer.

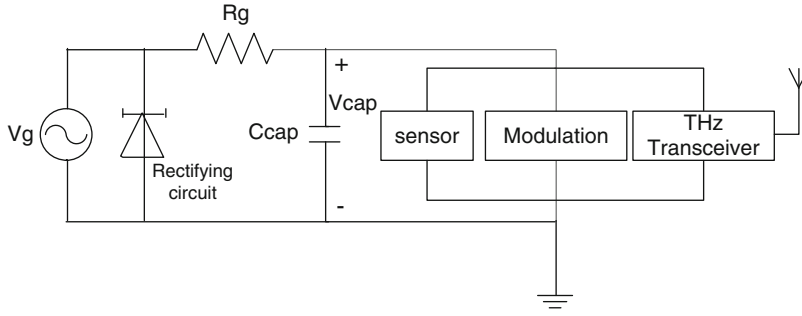
Foundations

One of the major bottlenecks in nanonetworks is the very limited energy that can be accessed by nanodevices. To achieve perpetual data transmission, it is required to investigate in-depth the relationship between energy harvesting and consumption and the underlying constraints in nanonetworks (Yao et al. 2015).

Energy Harvesting Modeling

In nanonetworks, the lifetime of each nanodevice is mainly dependent on the limited energy provided by the nanobatteries with a very small size. Hence, it is highly desirable that nanodevices are enabled to be self-powered without the use of nanobatteries. In the last decade, nanotechnology-enabled methods for converting mechanical energy into electrical energy by using the piezoelectric effect of zinc oxide nanogenerators have been explored (Wang and Song 2006). Due to the low energy consumption of nanodevices during the transmission of information in the THz band, especially over a short distance, the resultant energy harvested from the environment could be sufficient to power the nanodevices.

According to the simplified circuit model of piezoelectric nanogenerator as shown in Fig. 1, through several cycles of charging, the voltage of the nanocapacitor rises up, and the harvested energy is also stored in the nanocapacitor to power other modules in a nanodevice (Jornet and Akyildiz 2012). The maximum energy $E_{cap} - \max$ (n_{cyc}) stored in the nanocapacitors after a number of cycles n_{cyc} can be computed as a function of the total capacitance C_{cap} and voltage V_{cap} of the array of nanocapacitors:



Energy Modeling for Nanoscale Terahertz Communication, Fig. 1 Simplified circuit model of piezoelectric nanogenerator, which includes an array of ZnO nanowires

(i.e., electric source V_g), a rectifying circuit, and an ultra-nanocapacitor

$$E_{cap-\max}(n_{cyc}) = \frac{1}{2} C_{cap} (V_{cap}(n_{cyc}))^2. \quad (1)$$

For a particular nanodevice, its total capacitance C_{cap} and voltage V_{cap} are always predefined; then the required number of charging cycles n_{cyc} can be obtained when the energy consumption for transmission is determined, which will be comprehensively investigated in the next sections. Finally, the energy harvesting rate λ_{harv} in Joule/second at which the capacitors are charged can be obtained as follows (Jornet and Akyildiz 2011; Yao et al. 2015):

$$\lambda_{harv} = \frac{1}{t_{cyc}} \cdot \frac{\partial E_{cap}}{\partial n_{cyc}} = \frac{V_g \cdot \Delta Q}{t_{cyc}} \cdot \left(e^{\left(-\frac{\Delta Q \cdot n_{cyc}}{V_g C_{cap}}\right)} - e^{\left(-2\frac{\Delta Q \cdot n_{cyc}}{V_g C_{cap}}\right)} \right) \quad (2)$$

where ΔQ refers to the amount of electric charge obtained from one cycle, $\frac{\partial E_{cap}}{\partial n_{cyc}}$ refers to the energy increase of the nanocapacitors in each cycle, and t_{cyc} is the required time of one charging cycle, i.e., the time between two consecutive cycles. If the compress-release cycles are created by an artificially generated vibration source, the value of t_{cyc} corresponds to the inverse of the frequency of the vibration source. In this paper, the minimum value of n_{cyc} can be calculated to guarantee the required energy for transmission in the THz band.

Energy Consumption Modeling

Due to the constrained energy in nanodevices, short pulses cannot be emitted in a burst. Hence, new information encoding and modulation mechanisms for nanonetworks are required urgently. Furthermore, in order to take advantage of the large bandwidth of the THz band, pulse-based communication paradigm (such as TSOOK) has been recommended in (Jornet and Akyildiz 2011, 2014; Yao et al. 2015), which is based on the exchange of very short pulses spread in time. In detail, a logical “1” is transmitted by a short pulse (one hundred femtoseconds), and a logical “0” is transmitted as silence, i.e., nanodevices remain silent when a logical “0” is transmitted.

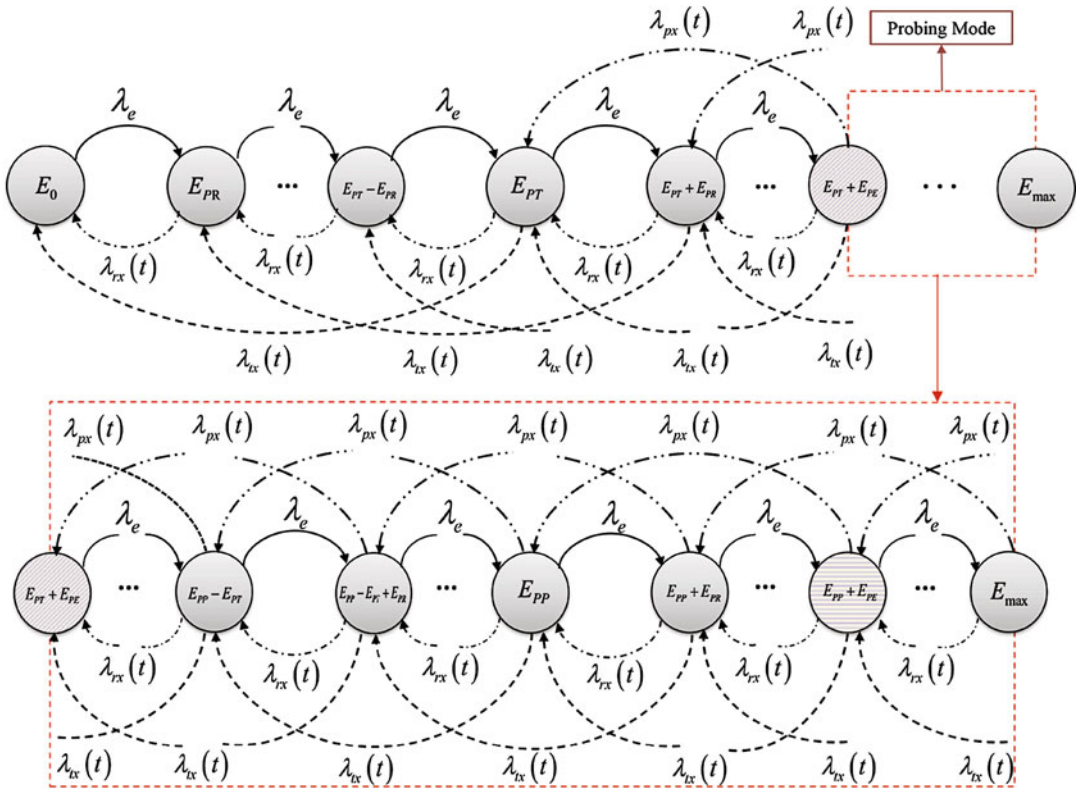
In order to evaluate the consumed energy in the transmission and reception of one packet, when the length of one packet is N_{bits}^y , the consumed energy in the transmission and reception of one pulse are E_{pul-t} and E_{pul-r} , respectively. Consequently, the consumed energy for transmitting and receiving one packet can be given by

$$E_{tx}^y = N_{bits}^y W_y E_{pul-t}, \quad (3)$$

$$E_{rx}^y = N_{bits}^y W_y E_{pul-r}, \quad (4)$$

where $y = p$ or $y = d$ stands for the calculation of one probing packet and one data packet, respectively. W_y refers to the coding weight, which is adaptive to different coding algorithms or parameter optimizations (Yao et al. 2015).

For simplicity, the general energy consumption model in nanonodes by considering the



Energy Modeling for Nanoscale Terahertz Communication, Fig. 2 The general process of energy state model in ECP mechanism, which considers the energy harvesting-consumption process based on extended Markov chain

energy harvesting-consumption process based on extended Markov chain approach (Jornet and Akyildiz 2012) is shown in Fig. 2, and the associated notations are listed in Table 1. Each state in the model corresponds to an energy state of nanonode. The red dashed box refers to a set of states that the nanonode works in the probing mode. The solid lines represent the energy harvesting process, while the other lines represent the energy consumption process. The short dashed lines represent the energy consumption processes of sending one data packet and receiving its feedback, which is denoted as λ_{tx} . The dash-dotted lines represent the energy consumption processes of receiving one data packet or one probing packet, which is denoted as λ_{rx} . The double-dash dotted lines represent the energy consumption processes of sending one probing packet and receiving

its feedback, which is denoted as λ_{px} . The detailed operations of the proposed model for electromagnetic nanonetworks are conducted as follows:

In order to increase the successful transmission probability of data packets and reduce the energy consumption of communication, this model divides all energy states into two parts as shown in Fig. 2. On the one hand, each nanonode will keep harvesting energy and enter the probing mode only as its energy value is beyond the probing mode energy threshold δ . On the other hand, in the probing mode, one probing packet will be transmitted to detect the channel condition at first. Data packet will be sent out only after the probing packet is transmitted successfully.

In general, two conditions in the energy state model are considered. The first condition is that nanonode only keeps harvesting energy without

Energy Modeling for Nanoscale Terahertz Communication, Table 1 Notations of the symbols in Fig. 2

Symbol	Description
$\{E\}$	The energy state when its energy value is E
E_0	Energy value of initial state
E_{PR}	Energy consumption of receiving one packet, which equals to the data receiving energy threshold τ
E_{PT}	Energy consumption of sending one packet and receiving its feedback, which equals to the data sending energy threshold γ
E_{PE}	Energy consumption of sending one probing packet and receiving its feedback
E_{max}	Energy saturated value
$E_{PT} + E_{PE}$	Energy threshold of the probing mode, δ
E_{PP}	An energy value between $E_{PT} + E_{PE}$ and E_{max}
λ_e	Energy harvesting rate
$\lambda_{tx}(t)$	The transmission rate of one data packet
$\lambda_{rx}(t)$	The reception rate of one data packet
$\lambda_{px}(t)$	The transmission rate of one probing packet

transmitting or receiving any packet from the initial energy state $\{E_0\}$ to the saturated energy state $\{E_{max}\}$. The second condition is that nanonode needs to transmit or receive data packets while harvesting energy. More specifically, there are two modes in the second condition as follows:

- Mode 1 (General Mode): From the beginning of initial energy state $\{E_0\}$, nanonode captures energy and receives packets (when necessary) rather than transmits any packet until it reaches the maximum energy E_{max} and enters the saturated energy state $\{E_{max}\}$. Moreover, the nanonode will not receive packets until the energy is beyond the data receiving energy threshold τ .
- Mode 2 (Probing Mode): While harvesting energy, nanonode needs to transmit data packets as well, if nanonode wants to transmit data in a certain state such as $\{E_{PP} + E_{PE}\}$. Firstly, its stored energy has to reach the energy threshold δ of probing mode. After sending out one probing packet, nanonode consumes the energy E_{PE} and enters into the energy state $\{E_{PP}\}$. According to the received feedback from the receiver after the probing packet transmitted, there are two cases:
 1. Case 1: If the feedback of probing packet is a negative acknowledgment

(NAK) or timeout, i.e., the channel is in bad condition, and nanonode needs to retransmit the probing packet. It is worth to note that nanonode has to determine whether its current energy reaches the energy threshold δ before detecting the channel by using the probing packet. As the energy is sufficient for channel detection, the probing packet will be sent out. The probing process continues until an ACK is received or its maximum number of retransmission is reached. Otherwise, the nanonode needs to harvest energy until the energy threshold δ , and then channel detection is performed again.

2. Case 2: If the feedback of probing packet is an active acknowledgment (ACK) and the energy of nanonode is beyond the energy threshold γ , then one data packet will be transmitted. This process will consume the energy E_{PT} , and let the nanonode move to the state $\{E_{PP} - E_{PT}\}$. For the feedback of the data packet, if it is an ACK, i.e., the data packet is transmitted successfully, the nanonode can start a new data transmission or end. If it is a NAK, the nanonode will enter into the probing mode again, until the data packet is transmitted successfully or its maximum number of retransmission is reached.

Key Applications

Energy modeling for nanoscale terahertz communication is fundamental for electromagnetic networks in the microscale or nanoscale applications where the energy is extremely limited or constrained.

Cross-References

- [Nanonetworks](#)
- [Nanoscale Terahertz Communications](#)
- [Wireless Sensor Networks](#)

References

- Akyildiz IF, Jornet JM (2010) The Internet of nano-things. *IEEE Wirel Commun* 17(6):58–63
- Jornet JM, Akyildiz IF (2011) Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the terahertz band. *IEEE Trans Wirel Commun* 10(10):3211–3221
- Jornet J, Akyildiz I (2012) Joint energy harvesting and communication analysis for perpetual wireless nanosensor networks in the terahertz band. *IEEE Trans Nanotechnol* 11(3):570–580
- Jornet J, Akyildiz I (2014) Femtosecond-long pulse-based modulation for terahertz band communication in nanonetworks. *IEEE Trans Commun* 62(5):1742–1754
- Krishnaswamy D, Helmy A, Wentzloff D (2010) Application of nanotechnologies in communications. *IEEE Commun Mag* 48(6):110–111
- Sudevalayam S, Kulkarni P (2011) Energy harvesting sensor nodes: survey and implications. *IEEE Commun Surv Tutor* 13(3):443–461
- Wang ZL (2008) Towards self-powered nanosystems: from nanogenerators to nanoelectronics. *Adv Funct Mater* 18(22):3553–3567
- Wang ZL, Song J (2006) Piezoelectric nanogenerators based on zinc oxide nanowire arrays. *Science* 312(5771):242–246
- Yao X-W, Wang W-L, Yang S-H (2015) Joint parameter optimization for perpetual nanonetworks and maximum network capacity. *IEEE Trans Mol Biol Multi-Scale Commun* 1(4):321–330
- Yao X-W, Wang C-C, Wang W-L, Jornet JM (2018) On the achievable throughput of energy-harvesting nanonetworks in the terahertz band. *IEEE Sensors J* 18(2):902–912

Energy Saving

- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)

Energy Saving of Cellular Networks

- [Energy Efficiency of Cellular Networks](#)

Energy Scavenging Technology

- [Energy Harvesting Technologies in Wireless Sensor Networks](#)

Energy-Aware and Throughput-Improved Route Selection for Wireless Multi-hop Networks

Yifei Wei

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Synonyms

[Routing protocol](#)

Definitions

As a means of providing instant networking to a group of devices which may not be within the

transmission range of one another without the support of any infrastructure, wireless multi-hop self-organizing networks are becoming increasingly popular, such as mobile ad hoc network (MANET), wireless sensor networks (WSNs), and wireless mesh networks (WMNs). Wireless multi-hop transmission is a promising wireless solution for numerous applications Ian et al. (2005), e.g., wireless sensor networking, broadband home networking, wireless campus networking, wireless enterprise networking, temporary emergency communications, intelligent transportation, etc. Despite the availability of many routing protocols for multi-hop wireless networks, the design of new routing metrics is still an active research topic in order to improve the network performance. The most frequently used metrics so far for routing protocols include hop count and link quality. However, they are far from satisfying the need of end users because finding an optimal route depends on the objectives and network characteristics.

Mobile devices are generally battery-powered and thus have limited energy resource that constrains the lifetime of the network. It is imperative that energy conservation is considered across all layers of the protocol stack in order to optimize the network energy consumption and prolong the lifetime of the network. Many theoretical studies show that energy consumption can be significantly reduced using energy-aware routing protocols compared to fixed-power minimum-hop routing protocols, and therefore the task of this subject is to design energy-aware and throughput-improved routing protocols for wireless multi-hop self-organizing networks.

Minimum Hop Count Routing Protocols

Most existing wireless routing protocols use hop count as the metric to choose the shortest path between source node and destination node. The traditional routing protocols such as dynamic

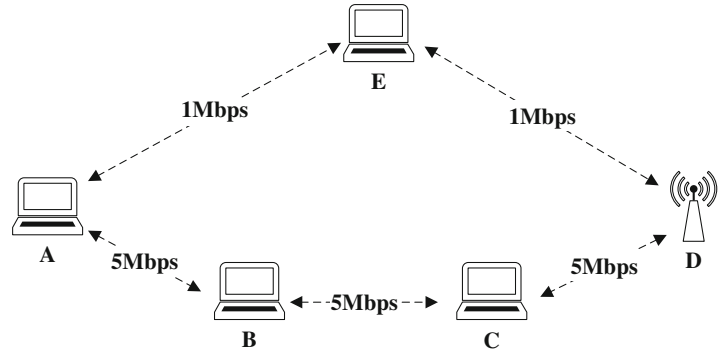
source routing (DSR) and ad hoc on-demand distance vector (AODV) are based on shortest path algorithms (or minimum hop count). In hop-count routing protocol such as AODV, source node sends a packet called routing frame in order to find the shortest path to the destination node. All the neighbor nodes of the source node will receive and forward this routing frame. Every routing frame has a hop counter which is called time to live (TTL) and has an initial positive integer. The TTL is decremented by one as packet travels one hop. When the TTL is reduced to zero, the routing frame will be discarded. After the destination node receives the routing frame forwarded by all its neighboring nodes, the destination node will pick up the routing frame with the highest TTL from the received routing frame to reply. The node that has forwarded the route request packet will form the shortest path from the source node to the destination node. After receiving the feedback, the source node will send information to the destination node according to this shortest path. This greedy hop-count routing protocol will lead every hops distance close to the nodes coverage radius. Also, one network with n connections, the routing frame will be re-transmitted n times if the source node does not know the way to destination node in the worst case. With the increase in nodes coverage radius, the number of connections between it and its neighbors will increase, which means that the times of retransmitting routing frame will increase. Even though, increase the radius of nodes will benefit the probability that finding a route from source node to destination node, too many times of retransmitting routing frame will increase the network overhead and possibility of routing loop which will influence other nodes extremely.

High-Throughput Routing Method

Minimum hop count routing protocols minimize the total number of transmissions required to send

Energy-Aware and Throughput-Improved Route Selection for Wireless Multi-hop Networks,

Fig. 1 Path selection options



a packet from source node to destination node. The minimum hop count routing metric is appropriate in single-rate wireless networks because every transmission consumes the same amount of resources. However, in multi-rate networks supporting the dynamic rate shifting scheme, the minimum hop metric has a tendency to select paths containing long-distance links that have low data rates and reduced reliability. Furthermore, the routing overhead will increase rapidly when the network topology changes frequently in mobile environment (Wei et al. 2010).

As is shown in Fig. 1, when the user equipment A want to communicate with the wireless access point D. According to the minimum hop count routing metric, the selected path will be $A \rightarrow E \rightarrow D$, which contains two long-distance links with data rates of 1 Mbps, respectively. Therefore, the whole path only provides the data rate of 500 Kbps and results in lowered network throughput and increased packet delay. To take advantage of multi-rate capabilities of wireless networks, we can select paths consisting of links with high data rates and better signal quality, but not necessarily with minimum number of hops. For example, when selecting the path of $A \rightarrow B \rightarrow C \rightarrow D$, the data rate can be more than 1 Mbps. Many works have been dedicated to providing high-throughput routing strategies. The previous work in Wei et al. (2011) optimizes the network throughput through relay selection under the finite-state Markov channels models. Fan (2004) proposes using MAC delay as a metric when choosing high-throughput routes in AODV.

The authors in Seok et al. (2003) introduced a multi-rate aware sublayer to find high data rate links.

Energy Hole Problem

Traditional routing protocols are local optimum and will produce the energy hole problem (Ren et al. 2016). There are two kinds of energy hole problems in multi-hop wireless networks: one happens in WSNs and another exists in wireless distributed networks.

The WSNs is known for monitoring the environment, where a sensor may act as an information detection device or a relay and a Sink node is used as the information collector. It is inconvenient or even infeasible to recharge or replace the batteries of sensors in WSNs. Traditional routing algorithm is based on the local optimization rather than global optimization. Therefore, sensors around the Sink node tend to run out of battery easily due to heavy traffic flow, which limits the network lifetime dramatically, and this problem is called the energy hole problem. The energy hole problem in WSNs can be solved by advanced routing protocol and power control (Guan et al. 2010).

In wireless distributed networks, nodes located near the center area of a region are more likely to be selected as a relay node. Therefore, the nodes in the center area usually relay too much data and run out of energy very quickly. For this reason, after several rounds of data

transmission, there would be an energy hole at the center of the convex region. Under the greedy routing algorithm, the energy hole will speed up the energy consumption of its nearby nodes, resulting in the energy hole expanding. The energy hole problem can also lead to the source node failure to reach the destination node, which is called routing failure. The energy hole problem in wireless distributed networks can be solved by topology control from a global perspective (Guo et al. 2014; Ao et al. 2014).

References

- Ao X, Yu FR, Jiang S, Guan Q, Leung VCM (2014) Distributed cooperative topology control for WANETs with opportunistic interference cancellation. *IEEE Trans Veh Tech* 63(2):789–801
- Fan Z (2004) High throughput reactive routing in multi-rate ad hoc networks. *Electron Lett* 40(25):1591–1592. <https://doi.org/10.1049/el:20046622>
- Guan Q, Yu FR, Jiang S, Wei G (2010) Prediction-based topology control and routing in cognitive radio mobile ad hoc networks. *IEEE Trans Veh Tech* 59(9):4443–4452
- Guo B, Guan Q, Yu FR, Jiang S, Leung VCM (2014) Energy-efficient topology control with selective diversity in cooperative wireless ad hoc networks: a game-theoretic approach. *IEEE Trans Wirel Commun* 13(11):6484–6495
- Ian FA, Wang X, Wang W (2005) Wireless mesh networks: A survey. *Computer Networks* 47(4):445–487
- Ren J, Zhang Y, Zhang K, Liu A, Chen J, Shen XS (2016) Lifetime and energy hole evolution analysis in data-gathering wireless sensor networks. *IEEE Trans Ind Inf* 12(2):788–800. <https://doi.org/10.1109/TII.2015.2411231>
- Seok Y, Park J, Choi Y (2003) Multi-rate aware routing protocol for mobile ad hoc networks. In: *The 57th IEEE semiannual vehicular technology conference, 2003. VTC 2003-Spring*, vol 3, pp 1749–1752. <https://doi.org/10.1109/VETECS.2003.1207123>
- Wei Y, Yu FR, Song M (2010) Distributed optimal relay selection in wireless cooperative networks with finite-state markov channels. *IEEE Trans Veh Technol* 59(5):2149–2158. <https://doi.org/10.1109/TVT.2010.2041803>
- Wei Y, Yu FR, Song M, Zhang Y (2011) Transmission control protocol throughput optimization in cooperative relaying networks through relay selection. *IET Commun* 5(16):2257–2265

Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks

Chengchao Liang

Department of System and Computer Engineering, Carleton University, Ottawa, ON, Canada

E

Synonyms

Traffic engineering; Software-defined wireless networking; In-network caching; Quality of experience

Definitions

Wireless Edge Caching (WEC) refers to that in-network caching is enabled at the edge of wireless networks such as access point and edge router. Software-defined wireless networking (SDWN) refers to that software-defined networking (SDN) is adopted in wireless networks to steer the traffic and in-network caching is a promising technology in the next-generation wireless networks. The quality of experience (QoE) of video streaming services mainly includes video resolutions, buffering delays, and stalling events. Then, it is important to consider the enhancement of the caching performance in wireless networks with considering QoE by utilizing the SDWN. First, bandwidth provisioning in SDWNs should be content-aware, which means it should assign network resources to users based on the caching status and the improvement of the QoE. Second, to enhance the hitting ratio of caches (utilization), caching strategies should be proactive according to the current traffic and resource status, behaviors of users, as well as the requirements of the QoE. Third, since video SDN flows from service providers usually have minimum require-

ments, the overall QoE performance of the network needs to be guaranteed.

Historical Background

Bandwidth provisioning (flow control) of SDWNs is studied in Farmanbar and Zhang (2014) and Dao et al. (2015) through traffic engineering. The authors of Farmanbar and Zhang (2014) propose a multipath traffic engineering formulation for downlink transmission considering both backhaul and radio access constraints. Moreover, the link buffer status is used as feedback to assist the adjustment of flow allocation. Based on Farmanbar and Zhang (2014), the authors of Dao et al. (2015) extend flow control of SDWNs to real-time video traffic specifically. This research proposes an online method to estimate the effective rate of video flows dynamically. Min flow rate maximization of SDWNs is investigated in Liao et al. (2014) with jointly considering flow control and physical layer interference management problem using weighted-minimum mean square error algorithm.

In the line of caching at mobile networks, the authors of El Bamby et al. (2014) formulate a delay minimization problem by optimally caching contents at the SBS. This research firstly clusters users with similar content and then deploys a reinforcement learning algorithm to optimize its caching strategy accordingly. Liu and Lau (2013) propose a scheme that BSs opportunistically employ cooperative multiple-input multiple-output (Coop-MIMO) transmission by caching a portion of files so that MIMO cooperation gain can be achieved without payload backhaul, while Dai and Yu (2014) and Tao et al. (2016) propose to utilize caching for releasing part of fronthaul in C-RAN. Unfortunately, the research does not take the real-time mobile network and traffic status into account. In Abboud et al. (2015), the proposed cache allocation policy considers both the backhaul consumption of small BSs (SBSs) and local storage constraints. The authors of Liang and Yu (2016) introduce in-network caching into SDWNs to reduce the latency of

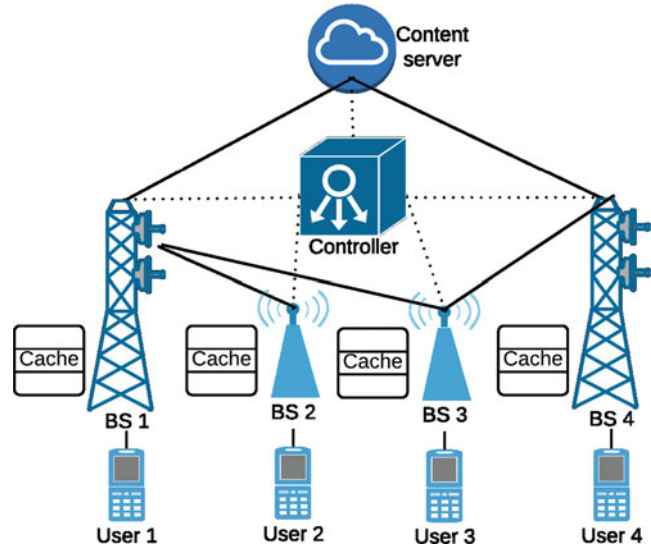
backhaul, but cache strategies are ignored in this study. Poularakis et al. (2015) introduces dynamic caching into BS selection from the aspect of energy saving. In Siris et al. (2014), dynamic caching is proposed to consider the mobility of users, but it focuses more on the cache resource and relationship between nodes. Network resources, such as the backhaul capacity and spectrum, are not discussed in this research.

Video quality adaptation with radio resource allocation in mobile networks (especially, long-term evolution (LTE)) is studied in a number of studies. The authors of Seufert et al. (2015) provide a comprehensive survey on quality of experience of HTTP adaptive streaming. Joint optimization of video streaming and in-network caching in HetNets can be traced back to Golrezaei et al. (2012). This research suggests that SBSs form a wireless distributed caching network that can efficiently transmit video files to users. Meanwhile, Ahlehagh and Dey (2014) take the backhaul and the radio resource into account to realize video-aware caching strategies in RANs for the assurance of maximizing the number of concurrent video sessions. The authors of Zhu et al. (2013) move the attention of in-network video caching to the core networks, instead of RANs, of LTE. To utilize new technologies in next-generation networks, Liu and Lau (2015) study the quality of video in next-generation networks. Pedersen and Dey (2016) conduct a comprehensive research that investigates opportunities and challenges of combining the advantages of adaptive bit rate and RAN caching to increase the video capacity and QoE of wireless networks. Another research combining the caching and video service is Yu et al. (2016), where collaborative caching is studied with jointly considering scalable video coding. However, RAN and overall video quality requirements are not considered.

Foundations

Generally speaking, SDWNs can enable the reduction of complexity and cost of networks, equip programmability into wireless networks, accelerate evolution of networks, and even further catalyze fundamental changes in the mobile ecosystem (Chen et al. 2015). The success of

**Enhancing QoE-Aware
Wireless Edge Caching
with Software-Defined
Wireless Networks, Fig. 1**
Network architecture of the
cache-enabled SDWN



E

the SDWN will depend critically on our ability to jointly provision the backhaul and radio access networks (RANs) (Liao et al. 2014).

WEC, as an extension of *in-network caching* that can efficiently reduce the duplicate content transmission (Ahlgren et al. 2012; Liang et al. 2015), has shown that access delays, traffic loads, and network costs can be potentially reduced by caching contents in wireless networks (Liu et al. 2016). To successfully combine caching and wireless networks, significant works (e.g., Li et al. 2015; Tao et al. 2016) have been done concerning utilizing and placing contents in the caches of base stations (BSs).

Thus, the network performance should be jointly optimized QoE of video streaming, bandwidth provisioning, and caching strategies in SDWNs with limited network resources and QoE requirements. To decrease the content delivery latency and improve the utilization of the network resources and caches, BSs are requested to retrieve data in advance dynamically according to the behaviors of users, the current traffic, and the resource status. To cope with the limited resources and the quality of service requirements, the caching problem is decoupled from the bandwidth provisioning problem by deploying the dual-decomposition method.

As shown in Fig. 1, a central SDN controller is deployed to control the network including

caching strategies and bandwidth provisioning. BSs connect to the CN through wired backhaul links with fixed capacities (e.g., bps). A content server is physically located at a core router (content delivery networks) in the CN or the source server. Moreover, as shown in Fig. 1, some BSs (e.g., SBSs) connect to the CN through MBSs. Without loss of generality, each BSs may have multiple links to be connected in the networks (e.g., SBS 3 in Fig. 1). To improve the caching performance, BSs can proactive cache contents that may be requested users. The probability that a cached content in BS j will be used by user i is assumed to be π_{ij} . In practice, π_{ij} depends on the user mobility pattern (Qiao et al. 2016) (current location, moving velocity and direction, and unpredictable factors) and the popularity of the content (Siris et al. 2014). The research about the behaviors of users has attracted great interests from both academia and industry (Yu and Leung 2001). Necessary information for processing prediction can come from real-time data (e.g., UEs and RANs) and historical data (e.g., users database). Fortunately, the improvement of machine learning and big data technologies can help the management and the control of mobile networks (He et al. 2016, 2017). By using those advanced techniques, the network is able to have a deeper view of the traffic and the coverage from the huge volume of historical data, which may

help the network to enhance the accuracy of the prediction.

By using the spare backhaul bandwidth to cache content to BSs before users actually served (or potentially continued being served) with those BSs (or actually request the content), BSs adaptively replace content in caches of BSs while optimizing the bandwidth provisioning and the video resolution selection. Moreover, caches and bandwidth are actually belonging to different layers of the network. Usually, flow bandwidth is restricted by physical resources (provisioned by media access control layer), and the video resolution depends on the achievable data rate. Joint optimization of video rate and flow bandwidth are studied in cross-layer design schemes (Yu and Krishnamurthy 2006). Differently, in-network cache is placed at the network layer or even higher layer. In addition, scheduling of the caching can be acting in a longer time period compared to flow scheduling.

Firstly, in the control plane, the SDN controller updates the network status according to feedback from BSs, then predict the users' behaviors to calculate $\{\pi_{ij}\}$. By using this information, the SDN controller evaluates and calculates the provisioned bandwidth and caching strategy for each flow i and BS j . After this, the SDN controller sends the control information to selected BSs and content source (the cloud center or the public networks). According to the control information, BSs and other network elements (e.g., potential routers) will update their SDN forwarding tables and perform resource allocation (e.g., cell association and scheduling). In the data plane, flows carrying contents requested by users can pass network elements selected by the control plane and to users.

Key Applications

The proposed scheme can be used to improve the QoE of multimedia transmission in the future wireless network. For example, this study can help the operator improve the QoE of video streaming in scenarios of cellular assisted-vehicle networks and high-speed trains in 5G.

Cross-References

- [Integrated System of Networking, Caching, and Computing](#)
- [Mobile Edge Caching in HetNets](#)
- [Wireless Edge Caching](#)

References

- Abboud A, Bastug E, Hamidouche K, Debbah M (2015) Distributed caching in 5G networks: an alternating direction method of multipliers approach. In: Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communication (SPAWC), Stockholm, Sweden pp 171–175
- Ahlegh H, Dey S (2014) Video-aware scheduling and caching in the radio access network. *IEEE/ACM Trans Netw* 22(5):1444–1462
- Ahlgren B, Dannewitz C, Imbrenda C, Kutscher D, Ohlman B (2012) A survey of information-centric networking. *IEEE Commun Mag* 50(7):26–36
- Chen T, Matinmikko M, Chen X, Zhou X, Ahokangas P (2015) Software defined mobile networks: concept, survey, and research directions. *IEEE Commun Mag* 53(11):126–133
- Dai B, Yu W (2014) Sparse beamforming and user-centric clustering for downlink cloud radio access network. *IEEE Access* 2:1326–1339
- Dao ND, Zhang H, Farmanbar H, Li X, Callard A (2015) Handling real-time video traffic in software-defined radio access networks. In: Proceedings of IEEE ICC workshops, London, UK pp 191–196
- El Bamby MS, Bennis M, Saad W, Latva-aho M (2014) Content-aware user clustering and caching in wireless small cell networks. In: International Symposium on Wireless Communication Systems (ISWCS), Barcelona, Spain pp 945–949
- Farmanbar H, Zhang H (2014) Traffic engineering for software-defined radio access networks. In: Proceedings of IEEE Network Operations & Management Symposium (NOMS), Krakow, Poland pp 1–7
- Golrezaei N, Shanmugam K, Dimakis AG, Molisch AF, Caire G (2012) Femtocaching: wireless video content delivery through distributed caching helpers. In: Proceedings of IEEE INFOCOM, Orlando, FL pp 1107–1115
- He Y, Yu FR, Zhao N, Yin H, Yao H, Qiu RC (2016) Big data analytics in mobile cellular networks. *IEEE Access* 4:1985–1996
- He Y, Liang C, Yu FR, Zhao N, Yin H (2017) Optimization of cache-enabled opportunistic interference alignment wireless networks: a big data deep reinforcement learning approach. In: Proceedings of IEEE ICC'17, Paris
- Li J, Chen Y, Lin Z, Chen W, Vucetic B, Hanzo L (2015) Distributed caching for data dissemination in

- the downlink of heterogeneous networks. *IEEE Trans Commun* 63(10):3553–3568
- Liang C, Yu FR (2016) Bandwidth provisioning in cache-enabled software-defined mobile networks: a robust optimization approach. In: *Proceedings of IEEE Vehicular Technology Conference (VTC) – Fall, Montréal, QC*
- Liang C, Yu FR, Zhang X (2015) Information-centric network function virtualization over 5G mobile wireless networks. *IEEE Netw* 29(3):68–74
- Liao WC, Hong M, Farmanbar H, Li X, Luo ZQ, Zhang H (2014) Min flow rate maximization for software defined radio access networks. *IEEE J Sel Areas Commun* 32(6):1282–1294
- Liu A, Lau VK (2013) Mixed-timescale precoding and cache control in cached MIMO interference network. *IEEE Trans Signal Process* 61(24):6320–6332
- Liu A, Lau VK (2015) Exploiting base station caching in MIMO cellular networks: opportunistic cooperation for video streaming. *IEEE Trans Signal Process* 63(1): 57–69
- Liu D, Chen B, Yang C, Molisch AF (2016) Caching at the wireless edge: design aspects, challenges, and future directions. *IEEE Commun Mag* 54(9):22–28
- Pedersen HA, Dey S (2016) Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing. *ACM/IEEE Trans Netw* 24(2):996–1010
- Poularakis K, Iosifidis G, Tassiulas L (2015) Joint caching and base station activation for green heterogeneous cellular networks. In: *Proceedings of IEEE International Conference on Communications (ICC), London, UK* pp 3364–3369
- Qiao J, He Y, Shen XS (2016) Proactive caching for mobile video streaming in millimeter wave 5G networks. *IEEE Trans Wirel Commun* 15(10): 7187–7198
- Seufert M, Egger S, Slanina M, Zinner T, Hoßfeld T, Tran-Gia P (2015) A survey on quality of experience of HTTP adaptive streaming. *IEEE Commun Surv Tutor* 17(1):469–492
- Siris VA, Vasilakos X, Polyzos G (2014) Efficient proactive caching for supporting seamless mobility. In: *Proceedings of IEEE 15th International Symposium on a World of Wireless, Mobile and Multimedia Network (WoWMoM), Sydney, Australia* pp 1–6
- Tao M, Chen E, Zhou H, Yu W (2016) Content-centric sparse multicast beamforming for cache-enabled cloud RAN. *IEEE Trans Wirel Commun* 15(9):6118
- Yu F, Krishnamurthy V (2006) Effective bandwidth of multimedia traffic in packet wireless CDMA networks with LMMSE receivers: a cross-layer perspective. *IEEE Trans Wirel Commun* 5(3):525–530
- Yu F, Leung VCM (2001) Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks. In: *Proceedings of IEEE INFOCOM'01, Anchorage*
- Yu R, Qin S, Bennis M, Chen X, Feng G, Han Z, Xue G (2016) Enhancing software-defined RAN with collaborative caching and scalable video coding. In: *Proceed-*

ing of IEEE International Conference Communications (ICC), Kuala Lumpur, Malaysia pp 1–6

Zhu J, He J, Zhou H, Zhao B (2013) Epcache: in-network video caching for LTE core networks. In: *Proceedings of International Conference Wireless Communications & Signal Processing (WCSP), Hangzhou, China* pp 1–6

Enzyme Reaction

► [Reaction-Diffusion Channels](#)

Equalization Techniques for Single-Carrier Modulations

Michael Rice

Department of Electrical and Computer Engineering, Brigham Young University, Provo, UT, USA

Synonyms

[Inter-symbol interference](#); [Linear modulation](#); [Maximum likelihood sequence detection](#)

Definitions

Linear modulations use pulses to transmit data over a channel. Transmission through a nonideal channel (say, from multipath propagation) causes temporal spreading of the pulses. The temporal spreading of the pulses produces a phenomenon known as inter-symbol interference: the detection instant for a given symbol contains contributions from adjacent symbols. Equalization refers to the category of signal processing techniques designed to mitigate inter-symbol interference.

Introduction

Linear, single-carrier modulations are described by the complex-valued, low-pass equivalent pulse

train (Proakis 2008).

$$s_c(t) = \sum_{k=0}^{K-1} a_k p_c(t - kT_s) \quad (1)$$

where a_k is the k -th complex-valued symbol drawn from an alphabet $\mathcal{C}$ with M members and with $E_s = E\{|a|^2\}$, $p_c(t)$ is a unit energy pulse shape, and T_s is the symbol time (i.e., a new M -ary symbol is transmitted every T_s sec). The number of symbols is K where K may be infinite. In most cases, the pulse shape $p_c(t)$ is designed to satisfy the Nyquist no-ISI theorem:

$$\int p(t)p(t - kT_s) dt = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \quad (2)$$

and is used to shape the power spectral density of $s_c(t)$.

When the pulse train is transmitted through a propagation medium described by the linear time-invariant system with impulse response $h_c(t)$, the pulse shape $p_c(t)$ may be altered by $h_c(t)$. The received signal is a pulse train based on the composite pulse shape $h(t) = p_c(t) * h_c(t)$. Including additive thermal noise, the received signal is modeled as

$$y_c(t) = \sum_{k=0}^{K-1} a_k h(t - kT_s) + z_c(t) \quad (3)$$

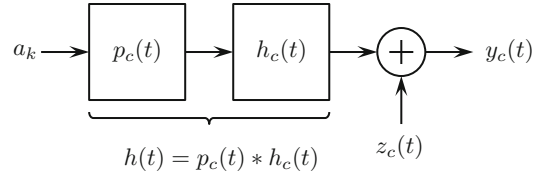
where $z_c(t)$ models the additive thermal noise and is a complex-valued zero-mean circularly symmetric normal random process with

$$E\{z_c(t + \tau) z_c^*(t)\} = 2N_0 \delta(\tau) \quad (4)$$

where $\delta(\tau)$ is the Dirac delta function. The situation is illustrated in Fig. 1. The problem is to estimate the sequence $a_0 \dots, a_{K-1}$ from $y_c(t)$.

If $h_c(t)$ is anything but a memoryless system, the composite pulse shape does not satisfy Eq. (2):

$$\int h(t)h(t - kT_s) dt \neq \begin{cases} 1 & k = 0 \\ 0 & k \neq 0. \end{cases} \quad (5)$$



Equalization Techniques for Single-Carrier Modulations, Fig. 1 The complex-valued, low-pass equivalent system considered for the equalizers

Because $h(t)$ does not satisfy Eq. (2), the usual matched filter detector outputs at $t = nT_s$ depend not only on the symbol a_n but also on temporally adjacent symbols. This dependency is a form of interference known as intersymbol interference (ISI). The goal of equalizers is to estimate the sequence $a_0 \dots, a_{K-1}$ from $y_c(t)$ in the presence of ISI.

Four broad approaches are described: maximum likelihood solutions in the section “[Maximum Likelihood Solutions](#),” filter-based equalizers in the section “[Filter-Based Equalizers](#),” adaptive equalizers in the section “[Adaptive Equalizers](#),” and blind equalizers in the section “[Blind Equalizers](#).”

Maximum Likelihood Solutions

The maximum likelihood (ML) solution applies to $y_c(t)$ a filter matched to $h(t)$ and uses T_s -spaced samples of the matched filter output with a sequence detector. The system that performs these functions is illustrated in Fig. 2. The output $r'_c(t)$ is given by

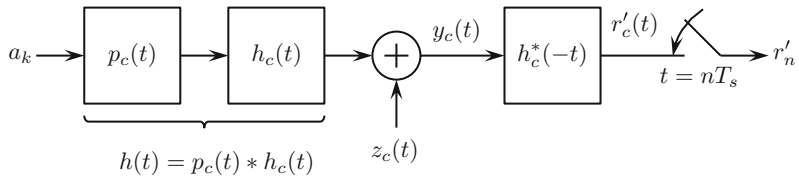
$$r'_c(t) = \sum_{k=0}^{K-1} a_k g(t - kT_s) + v_c(t) \quad (6)$$

where $g(t) = h(t) * h^*(-t)$ is the autocorrelation function of the composite pulse shape and $v_c(t)$ is a complex-valued wide-sense stationary circularly symmetric normal random process with autocorrelation function

$$E\{v_c(t + \tau) v_c^*(t)\} = 2N_0 g(\tau). \quad (7)$$

Equalization Techniques for Single-Carrier Modulations, Fig. 2

The complex-valued, low-pass equivalent system that represents the ML solution



It is customary to consider only pulse shapes and channels that produce an autocorrelation function $g(\tau)$ with support on $-LT_s \leq \tau \leq LT_s$ for some positive integer L . This consideration does not diminish the application of following results, because essentially all practical propagation channels are represented by linear systems with a finite-length impulse response.

The n -th T_s -spaced sample of $r'_c(t)$ is

$$r'_n \triangleq r'_c(nT_s) = \sum_{k=0}^{K-1} a_k g_{n-k} + v_n \quad (8)$$

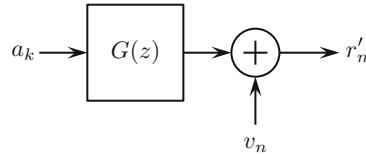
where $g_m \triangleq g(mT_s)$ and $v_n \triangleq v_c(nT_s)$. Equation (8) represents an entirely discrete-time system: the discrete-time inputs are the symbols and the discrete-time outputs are the samples r'_n . The *equivalent discrete-time system* is illustrated by the block diagram in Fig. 3. The discrete-time autocorrelation function g_m has support on $-L \leq m \leq L$. In the figure, $G(z)$ is the z -transform of g_m

$$G(z) = \sum_{m=-L}^L g_m z^{-m}. \quad (9)$$

The noise sample v_n is the n -th sample in a complex-valued wide-sense stationary circularly symmetric normal random process with zero mean and autocorrelation function

$$E\{v_n v_m^*\} = 2N_0 g_{n-m}. \quad (10)$$

The equivalent system in Fig. 3 is informally called the “Ungerboeck Observation Model” after the first person to analyze it (Ungerboeck 1974). In the Ungerboeck Observation Model, the sequence of r'_n forms the input to a ML sequence estimator based on the Viterbi algorithm. The Viterbi algorithm is based on a trellis with M^L



Equalization Techniques for Single-Carrier Modulations, Fig. 3 The equivalent discrete-time system corresponding to Fig. 2. This discrete-time model for the relationship between a_k and r'_n is informally known as the “Ungerboeck Observation Model”

states and uses the branch metric

$$J_U(a_n) = \text{Re} \left\{ a_n^* \left(2r'_n - g_0 a_n - 2 \sum_{m=1}^L g_m a_{n-m} \right) \right\} \quad (11)$$

for the symbol a_n connecting the state defined by $(a_{n-L}, \dots, a_{n-1})$ to the state defined by $(a_{n-L+1}, \dots, a_n)$.

In the Ungerboeck Observation Model (8), the noise samples accompanying the signal portion are correlated. It is usually preferred to perform sequence estimation in the presence of uncorrelated (white) noise samples. With uncorrelated noise samples, the branch metrics used by the Viterbi algorithm are more simple and the performance analysis is more straightforward. “Whitening” the noise sequence in Eq. (8) produces a discrete-time observation model informally known as the “Forney Observation Model,” after the first person to analyze it (Forney 1972).

The Forney Observation Model is derived from the Ungerboeck Observation Model as follows. The power spectral density of the noise term v_n in Eq. (8) is $2N_0 G(z)$. Noise whitening is based on the spectral factorization of $G(z)$. Because g_m is an autocorrelation function, it possesses conjugate symmetry: $g_{-m} = g_m^*$. Consequently, the $2L$ roots of the degree- $(2L+1)$

polynomial $G(z)$ defined in Eq. (9) also possess symmetry: $G(\rho) = 0 \Rightarrow G(1/\rho^*) = 0$. This means that each root inside the unit circle has a companion root outside the unit circle in the conjugate reciprocal position. Thus $G(z)$ may be written as

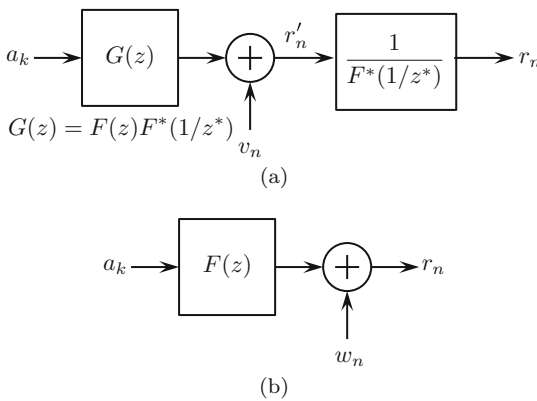
$$G(z) = F(z)F^*(1/z^*) \quad (12)$$

where $F(z)$ is the degree- $(L + 1)$ polynomial formed from the L roots of $G(z)$ inside the unit circle and $F^*(1/z^*)$ is the degree- $(L + 1)$ polynomial formed from the L roots of $G(z)$ outside the unit circle. Noise whitening is accomplished by passing r'_n through the anticausal IIR system $1/F^*(1/z^*)$ as illustrated in Fig. 4. The resulting output r_n is given by

$$r_n = \sum_{\ell=0}^L f_{\ell} a_{n-\ell} + w_n \quad (13)$$

where f_k for $k = 0, \dots, L$ is the k -th element of the inverse z -transform of $F(z)$, and w_n is a complex-valued wide-sense stationary circularly symmetric normal random process with zero mean and autocorrelation function

$$E\{w_n w_m^*\} = 2N_0 \delta_{n-m} \quad (14)$$



Equalization Techniques for Single-Carrier Modulations, Fig. 4 The equivalent discrete-time system known as the “Forney Observation Model”: (a) The relationship between the Forney Observation Model and the Ungerboeck Observation Model (Fig. 3); (b) The “Forney Observation Model”

where δ_{n-m} is the Kronecker delta function. The sequence of r_n in Eq. (13) forms the input to a ML estimator using the Viterbi algorithm based on M^L states and using the branch metric

$$J_F(a_n) = \left| r_n - \sum_{\ell=0}^L f_{\ell} a_{n-\ell} \right|^2 \quad (15)$$

for the branch connecting the state defined by $(a_{n-L}, \dots, a_{n-1})$ to the state defined by $(a_{n-L+1}, \dots, a_n)$.

In summary, the maximum likelihood detector for ISI channels is a maximum likelihood *sequence* estimator, usually realized by the Viterbi algorithm. A straightforward application of the maximum likelihood principle produces the equivalent discrete-time system known as the “Ungerboeck Observation Model” given by Eq. (8) and illustrated in Fig. 3. The “whitened” version is called the “Forney Observation Model” given by Eq. (13) and illustrated in Fig. 4b. Both achieved the same probability of error performance (Ungerboeck 1974). The differences lie in where the computational complexity resides. The branch metrics in the “Ungerboeck Observation Model” (11) require more arithmetic operations than the branch metrics in the “Forney Observation Model” (15). On the other hand, the “Forney Observation Model” requires a spectral factorization and a whitening filter operation. Both operate on a trellis with M^L states. Because the size of the trellis grows exponentially with L , the length of the composite filter, the computational complexity can be prohibitive. For this reason, nonoptimal approaches based on FIR filters are popular.

Filter-Based Equalizers

Equalizers based on discrete-time filters are used when the length of the equivalent discrete-time filter is too long for the ML approach to be computationally practical. Historically, finite-length impulse response (FIR) discrete-time filters are used because, unlike infinite-length impulse

response (IIR) discrete-time filters, the filter operation may be pipelined for fast operation, an important consideration in high data-rate applications.

FIR equalizer filters may operate on the output r'_n of the Ungerboeck Observation Model, on the output r_n of the Forney observation model, or on the outputs of a filter matched to the undistorted pulse shape $p_c(t)$. To simplify the discussion in this entry, filter-based equalizers operating with the Forney Observation Model are the only ones described.

Linear Equalizers

The linear equalizer is an FIR filter operating on the outputs r_n of the Forney Observation Model. The arrangement is illustrated in Fig. 5 where $C(z)$ is the z -transform of the FIR equalizer filter. The output of the equalizer filter $\tilde{a}_n$ is the input to a symbol-by-symbol (SxS) detector to compute the symbol estimate $\hat{a}_n$.

The probability of error is the probability that $\hat{a}_n \neq a_n$.

The FIR equalizer filter is defined by the filter coefficients

$$c_{-L_1} \quad \cdots \quad c_0 \quad \cdots \quad c_{L_2} \quad (16)$$

whose corresponding z -transform is

$$C(z) = \sum_{n=-L_1}^{L_2} c_n z^{-n}. \quad (17)$$

This formulation is general in the sense that a purely causal FIR filter is the special case $L_1 = 0$ and a symmetric filter is the special case $L_1 = L_2$. Starting with Eq. (13), the equalizer filter output is

$$\tilde{a}_n = \sum_{i=-L_1}^{L_2+L} q_i a_{n-i} + \sum_{i=-L_1}^{L_2} c_i w_{n-i} \quad (18)$$

where q_i is the impulse response of the cascade of the channel filter and the equalizer filter:

$$q_i = \sum_{j=-L_1}^{L_2} c_j f_{i-j}, \quad -L_1 \leq i \leq L_2 + L. \quad (19)$$

The equalizer output may be expressed as

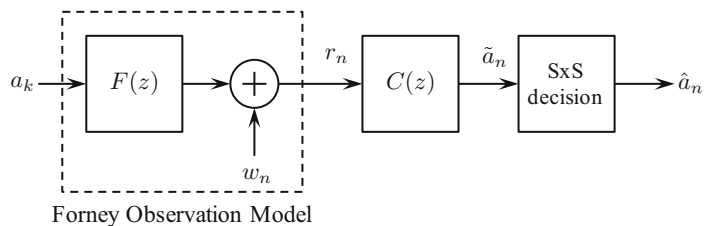
$$\tilde{a}_n = q_0 a_n + \sum_{\substack{i=-L_1 \\ i \neq 0}}^{L_2+L} q_i a_{n-i} + \sum_{i=-L_1}^{L_2} c_i w_{n-i}. \quad (20)$$

The first term on the right-hand-side of Eq. (20) is a scaled version of a_n and is the desired input to the SxS detector. If this were the only term in Eq. (20) and $q_0 = 1$, then the SxS detector would never make an error. The second term on the right-hand-side of Eq. (20) is the intersymbol-interference (ISI) present in the equalizer filter output. The goal of any good equalizer filter is make this term as small as possible. The third term on the right-hand-side of Eq. (20) is the contribution of the additive noise to the equalizer filter output. There are two popular criteria by which the equalizer filter coefficients are defined: The zero-forcing (ZF) criterion and the minimum mean-squared error (MMSE) criterion. The differences between them are how they incorporate the second and third terms on the right-hand-side of Eq. (20).

The ZF equalizer selects the filter coefficients to minimize the ISI term in Eq. (20). The original formulation of the problem (Lucky 1965) was based on the absolute value of the ISI:

Equalization Techniques for Single-Carrier Modulations, Fig. 5

The relationship between the FIR equalizer filter, represented by its z -transform $C(z)$, and the Forney Observation Model



$$\left| \sum_{\substack{i=-L_1 \\ i \neq 0}}^{L_2+L} q_i a_{n-i} \right| \leq A \left| \sum_{\substack{i=-L_1 \\ i \neq 0}}^{L_2+L} q_i \right| \leq A \sum_{\substack{i=-L_1 \\ i \neq 0}}^{L_2+L} |q_i| \quad (21)$$

where $A = \max_n \{a_n\}$. The object function to be minimized is

$$\sum_{\substack{i=-L_1 \\ i \neq 0}}^{L_2+L} |q_i| \quad (22)$$

which is called the maximum (or peak) distortion (Lucky 1965). For an infinitely long equalizer filter ($L_1 = L_2 = \infty$), the equalizer filter coefficients may be selected to produce $q_n = \delta_n$. For an FIR equalizer filter ($L_1 < \infty$, $L_2 < \infty$), the minimization produces $q_n = 0$ only for $-L_1 \leq n \leq L_2 + L$. This means the second term in Eq. (20) is not entirely eliminated. A second approach uses as the object function the least squares criterion (Hayes 1996).

$$|q_n - \delta_{n-n_0}|^2. \quad (23)$$

Choosing the equalizer filter coefficients to minimize (23) produces, for the equalizer filter, a filter that is the best least squares approximation to the inverse filter.

The shortcoming of the ZF equalizer is that the design criteria neglects the noise term in Eq. (20). The impact of this neglect is most dramatic when equalizing a channel whose discrete-time Fourier transform (DTFT) $|H(e^{j\omega})|$ displays one or more nulls. (The term “null” is used loosely here to mean not only the property that $|H(e^{j\omega})| = 0$ but also $|H(e^{j\omega})|$ is very small, say less than 0.1 times the maximum values of $|H(e^{j\omega})|$.) The DTFT of the ZF equalizer $C(e^{j\omega})$ is the inverse or approximate inverse of $H(e^{j\omega})$, i.e., $H(e^{j\omega})C(e^{j\omega}) \approx 1$. The action of $C(e^{j\omega})$ “equalizing” the amplitude variations in $H(e^{j\omega})$ is the source of the name “equalizer.” For channels with nulls, the

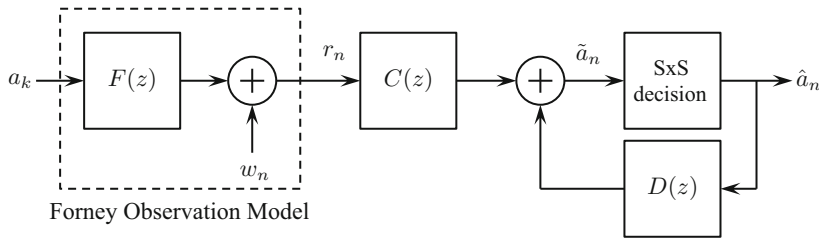
equalizer multiplies the spectral components corresponding to the channel nulls by large values. But the noise components corresponding to the frequencies of the nulls are also multiplied by the large values. The result is a noise term in Eq. (20) with very high variance. This property is often called “noise enhancement” and represents one of the challenging trade-offs with linear equalizers: ISI is reduced at the expense of reduced signal-to-noise ratio.

One way to address the shortcoming to the ZF equalizer is to adopt the mean-squared error criterion (Widrow 1966).

$$E \left\{ |a_n - \tilde{a}_n|^2 \right\}. \quad (24)$$

The solution for the equalizer filter coefficients is a form of the Wiener-Hopf equations and the resulting equalizer is known as the minimum mean-squared error (MMSE) equalizer. Because the mean-squared error criterion includes both the ISI and the noise terms in Eq. (20), the MMSE equalizer is able to balance the contributions of both ISI and noise enhancement in the solution. The result is less noise enhancement, but noise enhancement can still be a problem. In the limit as $N_0 \rightarrow 0$, the MMSE equalizer converges to the ZF equalizer.

The length of the ZF or MMSE equalizer $L_1 + L_2 + 1$ is an important consideration. The computational complexity increases with equalizer length, and this recommends using an equalizer as short as possible. In opposition, the ability of the equalizer to remove ISI increases with length, and this recommends using an equalizer as long as possible. Whereas increasing equalizer length improves performance, it is usually the case that increasing the equalizer length beyond a certain value returns diminishing improvements in performance. What this length is depends on the characteristics of the channel: its length L and the degree of amplitude variation with frequency. Experiments in Treichler et al. (1996) recommend an equalizer length 3–5 times L . But, short channels with deep spectral nulls require an equalizer much longer than this.



Equalization Techniques for Single-Carrier Modulations, Fig. 6 A block diagram of the decision-feedback equalizer based on the Forney Observation Model

E

Nonlinear Equalizers: The Decision-Feedback Equalizer

Both the ZF and MMSE equalizers suffer from a reduction in signal-to-noise ratio due to noise enhancement when the channel exhibits spectral nulls. In such circumstance, a nonlinear equalizer, known as the decision-feedback (DF) equalizer, exhibits better performance (Austin 1967).

The DF equalizer comprises two FIR filters arranged as shown in Fig. 6. The first filter is called the feedforward filter and is defined by the coefficients $c_{-L_1}, \dots, c_0$ with z -transform

$$C(z) = \sum_{n=-L_1}^0 c_n z^{-n}. \quad (25)$$

The second filter is called the feedback filter and is defined by the coefficients $d_1, \dots, d_{L_2}$ with z -transform

$$D(z) = \sum_{n=1}^{L_2} d_n z^{-n}. \quad (26)$$

This formulation follows the well-established convention of designating the feedforward filter as an anticausal filter and the feedback filter as a causal filter. The input to the SxS detector is

$$\tilde{a}_n = \sum_{i=-L_1}^0 c_i r_{n-i} + \sum_{j=1}^{L_2} d_j \hat{a}_{n-j}. \quad (27)$$

The first term on the right-hand-side of Eq. (27) represents the action of the feedforward filter that reduces, but does not eliminate, the ISI in r_n .

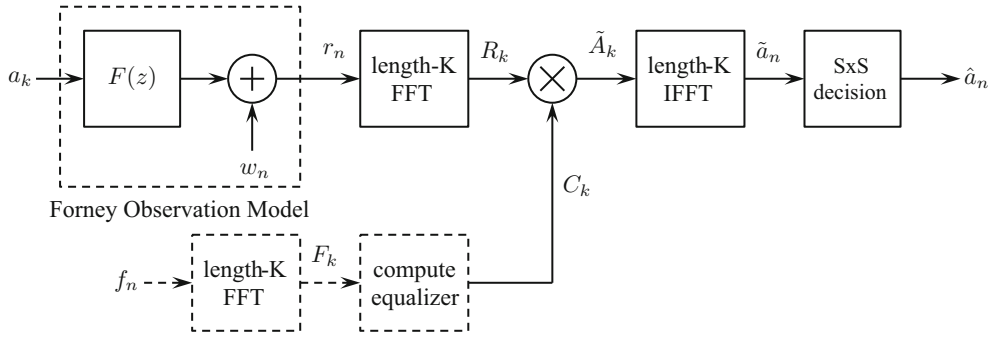
The second term on the right-hand-side of Eq. (27) reproduces the ISI remaining in the output of the feedforward filter and subtracts it from the output of the feedforward filter in the hopes of producing an ISI-free input to the SxS detector. Noise enhancement is greatly reduced because the feedback filter operates on symbol decisions which, if they are correct, are noise free. Either the ZF or MMSE criterion may be used to compute the coefficients of the feedforward filter. The MMSE is the most common. The feedback filter coefficients are set to the negatives of the causal part of $q_n = f_n * c_n$.

Choosing the equalizer length is a two-step process. The performance vs. length trade-off for feedforward equalizer follows that of the MMSE equalizer described above. The length of the feedback equalizer filter needs to be at most L .

Frequency-Domain Equalizers

Some channels require very long equalizers to achieve satisfactory performance; frequency-domain equalizers address the computational burden by leveraging the computational advantages of the FFT algorithm. The filtering operation required by the linear equalizer is accomplished by performing point-by-point multiplications of the discrete-Fourier transforms (DFT) of the received signal samples and the equalizer filter.

The frequency-domain equalizer was first described in Walzman and Schwartz (1973). A block diagram illustrating the idea is shown in Fig. 7. The k -th element of the length- K FFT of the received sequence r_n is



Equalization Techniques for Single-Carrier Modulations, Fig. 7 A block diagram illustrating the frequency-domain equalizer based on the Forney Observation Model

$$R_k = \sum_{n=0}^{K-1} r_n e^{-j2\pi nk/K}, \quad k = 0, \dots, K-1. \quad (28)$$

The relationship between the f_n and F_k is identical. Equalization is accomplished in the frequency domain by performing a point-by-point multiplication of R_k and the equalizer C_k . An inverse-FFT is applied to the product to return to the sample domain. The frequency-domain equalizer may be

$$C_k = \begin{cases} \frac{1}{F_k} & \text{ZF equalizer} \\ \frac{F_k^*}{|F_k|^2 + 2N_0/E_s} & \text{MMSE equalizer,} \end{cases} \quad k = 0, \dots, K-1. \quad (29)$$

Point-by-point multiplication in the DFT domain is equivalent to circular convolution/deconvolution in the sample domain. But the ISI occurs through linear convolution in continuous-time domain, cf. Figure 1. To match the two, a *cyclic prefix* is used at the transmitter to convert linear convolution to circular convolution in preparation for frequency-domain equalization.

The operations representing the computations of the channel F_k and the equalizer C_k are shown in dashed lines in Fig. 7. This is because for static channels, the FFT to produce F_k and the computations to produce C_k need to be performed only once, or infrequently in the case of slowly varying channels.

Summary

In summary, linear equalizers, based on equalizer filters optimizing the ZF or MMSE criteria, are low-complexity alternatives to the maximum likelihood sequence estimator. These equalizers comprise simple FIR filters followed by a symbol-by-symbol detector. For a channel with one or more spectral nulls, noise is the limiting factor in equalizer performance. Noise enhancement is reduced by using a decision-feedback equalizer. For long channels (large L), the length of the equalizer filter required to achieve acceptable performance can be prohibitively large. In this case, frequency domain equalization techniques can be used. Frequency-domain techniques require the insertion of a cyclic prefix to convert linear convolution into circular convolution and can be limited by noise enhancement.

Adaptive Equalizers

The equalizers described in the previous two sections relied on knowledge of the equivalent discrete-time channel relating the transmitted symbols to the received samples. In practice, this knowledge is obtained by estimating the equivalent discrete-time channel from the received samples corresponding to a known sequence symbol inserted into the data stream by the transmitter. The equalizer filter coefficients are computed from the channel estimate as the solution to matrix/vector equations. The

dimensions of the matrix and vector are determined by the equalizer length. For long equalizers, computing the solution can be prohibitively complex.

The alternative is to use an adaptive filter (Haykin 2013). Equalizers based on adaptive filters are called adaptive or automatic equalizers. The structure of the adaptive ZF and MMSE equalizers is identical to that shown in Fig. 5 except $C(z)$ is replaced by an adaptive filter. The structure of the adaptive DF equalizer is identical to that shown in Fig. 6 except $C(z)$ and $D(z)$ are replaced by adaptive filters.

An adaptive filter is one where the filter coefficients are recursively updated based on the error signal $e_n = a_n - \tilde{a}_n$. The LMS algorithm for the ZF filter coefficient updates (Haykin 2013) is

$$\begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_{n+1} = \begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_n + \mu e_n \begin{bmatrix} a_{n+L_1}^* \\ \vdots \\ a_n^* \\ \vdots \\ a_{n-L_2}^* \end{bmatrix} \quad (30)$$

where the subscript on the vector of equalizer filter coefficients indicates the time index that the filter coefficients are used. The recursion converges to the correct solution if the step size $\mu > 0$ is sufficiently small. The LMS algorithm for the MMSE equalizer filter coefficient updates is

$$\begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_{n+1} = \begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_n + \mu e_n \begin{bmatrix} r_{n+L_1}^* \\ \vdots \\ r_n^* \\ \vdots \\ r_{n-L_2}^* \end{bmatrix} \quad (31)$$

and the LMS algorithm for the DF equalizer filter coefficients is

$$\begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_{n+1} = \begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_n + \mu e_n \begin{bmatrix} r_{n+L_1}^* \\ \vdots \\ r_n^* \\ \vdots \\ a_{n-L_2}^* \end{bmatrix} \quad (32)$$

Recursions for the MMSE frequency domain equalizer filter coefficients C_k are described in Walzman and Schwartz (1973).

The advantages of the ZF, MMSE, and DF adaptive equalizers is that they do not have to “know” the channel f_n and do not have to compute the solutions to a potentially high-dimensional matrix vector equation. But, adaptive equalizers do need to be “trained”: a sufficiently long sequence of known symbols, often called a “training sequence,” is used as the known symbols in Eqs. (30), (31), and (32). The length of the training sequence must be on the order of the convergence time of the equalizer. The convergence time is determined by the step size μ and by the deterministic autocorrelation function of the channel f_n . Once the adaptive filter has converged, operation may shift to a decision directed mode where a_i is replaced by $\hat{a}_i$ in Eqs. (30), (31), and (32). This decision-directed mode allows the adaptive filter to track small, slow changes in the channel.

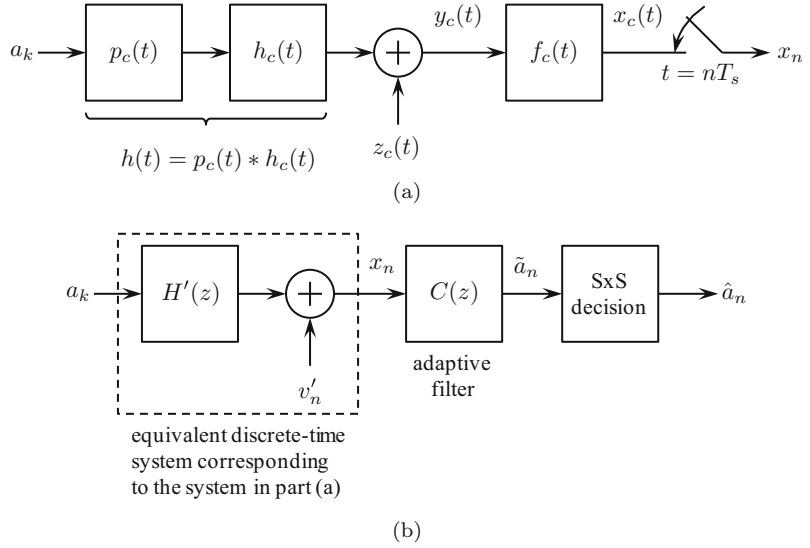
Convergence time can be reduced by using a more sophisticated recursion, such as the recursive least squares (RLS) algorithm Haykin (2013). Improved convergence time is achieved at the expense of increased computational complexity required by the recursive updates.

Blind Equalizers

Blind equalizers are based on adaptive filters that do not need to be trained. The adaptive filters use LMS-type updates based on a well-chosen non-linear property of the data symbols. The adaptive filter is able to converge to an equalizer filter that

Equalization Techniques for Single-Carrier Modulations, Fig. 8

Block diagrams of the system used to describe blind equalizers: (a) the continuous-time system of Fig. 1 using a receive filter with impulse response $f_c(t)$ whose output is sampled at T_s -spaced intervals; (b) the equivalent discrete-time system of corresponding to (a) together with the adaptive filter that forms the blind equalizer



gives good performance even without knowledge of the channel or the data symbols.

Blind equalizers are used in situations where the channel is unknown. Consequently, the Forney Observation Model (which requires knowledge of the channel) does not apply here. In its place, we consider the systems illustrated in Fig. 8. The block diagram in Fig. 8a starts with the system in Fig. 1 and adds a continuous-time receiver filter with impulse response $f_c(t)$. The receiver filter may be an ideal low-pass filter or a filter matched to the pulse shape. The output of the receiver filter is sampled at T_s -spaced intervals to produce the sequence x_n . The system in Fig. 8b illustrates the equivalent discrete-time system corresponding to Fig. 8a and the position of the adaptive filter that forms the blind equalizer. The equivalent discrete-time system comprises a filter with samples $h'_n \triangleq h'_c(nT_s)$ where

$$h'_c(t) = p_c(t) * h_c(t) * f_c(t). \quad (33)$$

The noise samples v'_n are samples of the output of receive filter due to $z_c(t)$ and comprise a complex-valued wide-sense stationary circularly symmetric discrete-time normal random process with autocorrelation function is formed from T_s -spaced samples of the temporal autocorrelation

function of the receive filter. The output of the adaptive filter forms the input to the SxS detector.

The equalizer filter coefficients are given by Eq. (16) and the update is of the form

$$\begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_{n+1} = \begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_n - \mu \nabla_n J \quad (34)$$

where $\nabla_n J$ is the gradient vector of a scalar valued object function J with respect to the equalizer filter coefficients at time step n . This is the gradient descent method for minimization: the filter coefficient vector at time step $(n+1)$ moves in the opposite direction of the gradient of the object function evaluated at the coefficient vector at time step n . The step size is controlled by the scalar $\mu > 0$.

The two most-studied object functions are the generalized Sato function (Benveniste and Goursat 1984):

$$J = \frac{1}{2} E \left\{ |\tilde{a}_n - \gamma \operatorname{csgn}(\tilde{a}_n)|^2 \right\} \quad (35)$$

where $\gamma = E\{|a_n|^2\}/E\{|a_n|\}$ and

$$\text{csgn}(\tilde{a}_n) = \text{sign}(\text{Re}\{\tilde{a}_n\}) + j \text{sign}(\text{Im}\{\tilde{a}_n\}), \quad (36)$$

and Godard's constant modulus function Godard (1980)

$$J = \frac{1}{2} \text{E} \left\{ \left[|\tilde{a}_n|^2 - R_2 \right]^2 \right\} \quad (37)$$

where $R_2 = \text{E}\{|a_n|^4\}/\text{E}\{|a_n|^2\}$. Of the two, Godard's constant modulus function is the most widely used.

The blind equalizer based on Eq. (37) uses an adaptive filter with the updates of the form Eq. (34) with J given by Eq. (37). Computing the gradient ∇J and using a single point estimate of the expectation in Eq. (37), the adaptive filter update is

$$\begin{aligned} \begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_{n+1} &= \begin{bmatrix} c_{-L_1} \\ \vdots \\ c_0 \\ \vdots \\ c_{L_2} \end{bmatrix}_n \\ &+ \mu \tilde{a}_n \left(|\tilde{a}_n|^2 - R_2 \right) \begin{bmatrix} x_{n+L_1}^* \\ \vdots \\ x_n^* \\ \vdots \\ x_{n-L_2}^* \end{bmatrix} \end{aligned} \quad (38)$$

This blind equalizer update algorithm is known as the constant modulus algorithm.

Summary

Equalization is used to enable the detection of data symbols in the presence of ISI caused by nonideal propagation channels. The ML solution comprises a sequence estimator usually implemented by the Viterbi algorithm operating on a trellis where the number of states grows exponentially with channel length. For long channels, the computational complexity of the ML estimator is prohibitive, so low-complexity filter-based

equalizers are used. Linear equalizers are based on an FIR filter and a symbol-by-symbol detector. The computational complexity of the equalizer is modest (except for the need to compute the equalizer filter coefficients from the known or estimated channel), but performance suffers from noise enhancement. The decision-feedback equalizer reduces the noise enhancement problem. The computational complexity of implementing long equalizer filters can be reduced by performing equalization in the DFT domain. Because DFT-domain multiplication corresponds to circular convolution, a cyclic prefix must be inserted at the transmitter to convert linear convolution to circular convolution. The computational complexity of computing the equalizer filter coefficients can be reduced by using adaptive filters: the equalizer filters converge to the true equalizer filter when operating on the samples corresponding to a sequence of known training symbols. Adaptive versions of the zero-forcing, MMSE, and decision-feedback equalizers are straightforward to implement. The need for training symbols can be eliminated using blind equalizers.

Cross-References

- [Fundamental Properties of Wireless Relays and Their Channel Estimation](#)
- [Massive MIMO Channel Estimation](#)
- [Millimeter Wave Channel Measure](#)
- [Millimeter Wave Channel Modeling](#)

References

- Austin M (1967) Decision feedback equalization for digital communication over dispersive channels. Technical report no. 437. MIT Lincoln Laboratory, Lexington
- Benveniste A, Goursat M (1984) Blind equalizers. *IEEE Trans Commun* 32(8):871–883
- Forney GD (1972) Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. *IEEE Trans Inf Theory* 8(3):363–378
- Godard D (1980) Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *IEEE Trans Commun* 28(11):1867–1875

- Hayes M (1996) Statistical digital signal processing and modeling. Wiley, Hoboken
- Haykin S (2013) Adaptive filter theory, 5th edn. Pearson Prentice-Hall, Upper Saddle River
- Lucky R (1965) Automatic equalization for digital communication. *Bell Syst Tech J* 44(4):547–588
- Proakis J (2008) Digital Communications, 5th edn. McGraw-Hill, New York
- Treichler J, Fijalkow I, Johnson C (1996) Fractionally spaced equalizers: how long should they really be? *IEEE Signal Process Mag* 13(3):65–81
- Ungerboeck G (1974) Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems. *IEEE Trans Commun* 22(5):624–636
- Walzman T, Schwartz M (1973) Automatic equalization using the discrete frequency domain. *IEEE Trans Inf Theory* 19(1):59–68
- Widrow B (1966) Adaptive filters I: fundamentals. Technical report no. 6764-6. Stanford Electronics Laboratory, Stanford University, Stanford

ETSI ITS-G5 Standard

► Wireless Access in Vehicular Environment

EV Charging Scheduling

Kuang-Hao Liu
Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

Synonyms

► Charging navigation

Definitions

Determine when the electric vehicles (EVs) should visit which charging station during the trip to achieve certain objectives, e.g., minimizing the waiting time or charging cost, maximizing the profit of charging network, load balancing of the electricity grid, and so on.

Historical Background

Charging scheduling plays an important role in enabling the intelligent transportation system with EVs. Unlike vehicles powered by gas or diesel, EVs are powered by electricity from in-car energy storage such as batteries or a fuel cell. Due to the limited capacity of in-car energy storage, EVs need to be charged more frequently than conventional cars. Frequent charging not only takes time (including the time to reach a charging station and that to complete charging) but also reduces the battery lifetime. Besides, electricity price varies in time and locations. An EV driver might choose to charge at private charging facilities such as home or office parking lot. There is also a need to charging at public charging stations, particularly for long trip and heavy-duty cars. Compared to the former case, a public charging station may be occupied by many drivers during a certain period, leading to long waiting time for charging. Hence, the charging decision of individual EV needs to be properly planned for the best driving experience and efficient utilization of electricity.

EV charging scheduling shares the common goal as route planning (Eisner et al. 2011), a function already available in many personal navigation devices. Both attempt to assist the EV drivers to reduce their traveling time. In essential, route planning finds the optimal end-to-end path for the individual driver based on some given information, such as departure and destination location, route preference, and so on. Charging scheduling guides the driver on when and where to charge during the trip. To this end, real-time status about the availability of charging stations and instantaneous traffic conditions must be taken into account in charging scheduling. While route planning can also respond to the traffic conditions, it is performed to minimize the traveling time of individual vehicles. Differently, charging scheduling is made considering the entire vehicular networks and can be used to adjust the load of electricity grid. As a result, charging scheduling has a broader impact to the entire transportation network and electricity power system.

In general, effective charging scheduling for EVs is difficult. For some scenarios, such as auto shipping and public transportation, the route can be known in a priori, but this may not be the case for personal travels or daily commuting that involve rapid changes on road conditions and per-driver preference. The random movement of EVs also causes uncertainties in both time and space. Additionally, the status of charging stations, such as availability and waiting lines, varies at different times and locations.

To support charging scheduling, real-time and reliable bidirectional communications between EVs and the control center that derives charging scheduling based on the collected information from EVs are of paramount importance. Considering the large number of EVs in the future, massive connectivity with high mobility is expected that challenges the existing wireless networks such as mobile cellular networks and wireless local area networks (WLAN) as both are not designed to serve EV-related communications. In the USA, the technology called dedicated short-range communications (DSRC) (Kenney 2011) or known as wireless access in vehicular environments (WAVE) has been standardized to support vehicular-to-vehicular (V2V) and vehicular-to-roadside (V2R) communications. However, existing DSRC has a bandwidth of 10 MHz per channel only and the achievable data rate up to 27 Mbps. On the other hand, fourth-generation cellular network or known as Long-Term Evolution (LTE) has a wider bandwidth and much higher data rate. Despite 4G-LTE will be enhanced in its future release to support massive connectivity with ultralow latency, it operates in the licensed band leading to a higher operational cost than DSRC. The debate over DSRC versus LTE remains open at both the regulation and commercialization level.

Apart from the underlying communication networks, efficient algorithms for EV charging scheduling serve as the key enabler for connected EVs. Wang et al. (2016a) provide a comprehensive survey on EV charging scheduling. Existing EV charging scheduling can be further classified

in two categories depending on the energy source of the charging stations. In the first category, charging stations are hooked up to the electrical grid, e.g., Tesla Supercharger stations (Tesla Motors 2017). Since the electrical grid is designed to be always on, EVs can be charged immediately at an available charging station. In this context, the EV charging scheduling can be designed, for example, to navigate EVs to the charging station that minimizes the time cost for charging, including the travel time to the charging station, the waiting time for an available outlet in the charging station, and the actual charging time, as studied by Wei et al. (2015). Meanwhile, the residual electricity of the EV needs to be taken into account when determining the route to an appropriate charging station. Such a problem has been shown by Zhu et al. (2014) as an NP-complete problem, and thus heuristic or approximated algorithms must be developed to quickly derive the scheduling results. To make electricity utilization more efficient, EVs may not only draw energy from the grid but also feed energy back to the grid, resulting in a more sophisticated design for two-way charging scheduling. As revealed by Zeng et al. (2013), two-way charging scheduling can increase the utilization of charging stations, reduce the charging delay, and improve the peak-to-average ratio of electric load.

The second category of EV charging scheduling considers off-grid charging stations that rely on renewable energy such as solar and wind. As a clean energy source, renewable energy produces much less pollution to the environment and thus promises a sustainable transportation system. There have been some initiatives being taken, e.g., Driving on Sunshine program (Charge Across Town 2018) in the San Francisco bay area and California, that demonstrate the great potential of off-grid charging stations. However, renewable energy is intermittent and unpredictable in its arrival pattern that challenges EV charging scheduling. In addition, renewable energy alone may not offer enough energy to fulfill the charging need in the public area. An alternative is thus to jointly use grid

power and renewable energy at the charging stations. In this case, EV charging scheduling needs to adapt to the intermittence of the renewable energy, the uncertain vehicle arrivals, and the variation of the grid power price. Such a problem can be formulated as a stochastic optimization problem under the framework of Markov decision process (MDP) (Zhang et al. 2014). The similar problem aiming to reduce the charging time is also studied by Wang et al. (2016b), which develops a two-stage EV charging mechanism to determine the optimal energy generation and charging rate of each EV.

Key Applications

Charging scheduling is essential to intelligent transportation system with EVs. Drivers use the scheduling information provided by a control center to decide when and where to charge during the trip in order to save traveling time and cost. The charging scheduling policy can incorporate the electricity price and load of charging stations so as to maximize the efficiency of entire charging network or the profit of charging service providers. The electricity grid can also leverage charging scheduling for load balancing and possibly optimizing the integration of renewable energy into the power system.

Cross-References

- [Charging Navigation](#)
- [Vehicle-to-Vehicle Communications](#)
- [Vehicular Networks](#)

References

- Charge Across Town (2018) Driving on sunshine. <http://chargeacrosstown.org/driving-on-sunshine/>
- Eisner J, Funke S, Storand S (2011) Optimal route planning for electric vehicles in large network. In: Proceedings of twenty-fifth AAAI conference on artificial intelligence

- Kenney JB (2011) Dedicated short-range communications (DSRC) standards in the united states. *Proc IEEE* 99(7):1162–1182
- Tesla Motors (2017) Super charger stations. <https://www.tesla.com/supercharger>
- Wang Q, Liu X, Du J, Kong F (2016a) Smart charging for electric vehicles: a survey from the algorithmic perspective. *IEEE Commun Surv Tutor* 18(2): 1500–1517
- Wang R, Wang P, Xiao G (2016b) Two-stage mechanism for massive electric vehicle charging involving renewable energy. *IEEE Trans Veh Technol* 65(6):4159–4171
- Wei Z, He J, Xing M, Cai L (2015) Utility maximization for electric vehicle charging with admission control and scheduling. In: Proceedings of IEEE international conference on communications (ICC)
- Zeng M, Leng S, Zhang Y (2013) Power charging and discharging scheduling for V2G networks in the smart grid. In: Proceedings of IEEE international conference on communications (ICC)
- Zhang T, Chen W, Han Z, Cao Z (2014) Charging scheduling of electric vehicles with local renewable energy under uncertain electric vehicle arrival and grid power price. *IEEE Trans Veh Technol* 63(6):2600–2612
- Zhu M et al (2014) The charging-scheduling problem for electric vehicle networks. In: Proceedings of IEEE wireless communications and networking conference (WCNC)

Event Detection and Reporting

- [Data Gathering in Wireless Sensor Networks](#)

Evolution of Vehicular Networks in the Transition from 3G to 4G Systems

Francesco Chiti
Department of Information Engineering,
University of Florence, Florence, Italy

Synonyms

[Cellular technologies](#); [Device-to-device communications](#); [Spontaneous networking](#); [Vehicular ad hoc networks](#)

Definitions

The role of cellular technologies in enabling ad hoc vehicular networks.

Introduction

The actual trends in wireless networking are (i) to extend services offered by traditional providers with unlimited peer-to-peer capabilities and (ii) to dynamically integrate available communications segments to exploit all communication opportunities. The latter feature is inherently dictated by the correlations among data flows which in turn are due to wide sense *social* relationship, i.e., the requirements arising when establishing a *group* of devices.

A challenging domain for the integration and evolution of heterogeneous information and communication technologies (ICTs) is represented by the automotive industry. Indeed, in order to provide drivers with a more pleasant and safer driving experience, vehicles are gradually endowed with sensing, actuation, computing, user interaction, and communications capabilities, thus becoming *intelligent nodes* capable of interconnecting, cooperating, and even autonomously adapting to the surrounding environment. Since drivers usually spend a non-negligible fraction of time in vehicles and their behaviour may have a strong impact at a collective scale (e.g., pollution and noise in a city), the adoption of advanced ICT approaches for achieving a more safe, green, and effective mobility management is still an open research issue (Gerla and Kleinrock 2011).

The familiarity with the *social networking* paradigm makes people used to share information online with (in)direct *contacts* and use it in their everyday life depending on the degree of trust. Moreover, the widespread adoption of existing social-based traffic and navigation applications paves the way for introducing novel vehicular services and applications that are typically location-based as well as community driven. To this purpose, wireless communications

may enable the effective exchange of *mobile* information among vehicles in a participatory way to enhance open information sharing and knowledge exchange processes. Despite 3G systems have been capable to support *generic* and *individual* terminal mobility, recently, investigation activities have been started on the possible use of 4G Long Term Evolution (LTE) standard to ensure connectivity between vehicles, roadside infrastructure, and the people inside and around the so-called connected car (3rd Generation Partnership Project – Technical Specification Group Services and System Aspects 2015). The definition of a generic interface, namely, V2X, between a vehicle and a generic device has been addressed, generalizing the existing communication interfaces as V2V (between vehicles), V2P (between a vehicle and a device carried by a generic individual), and V2I (between a vehicle and a roadside unit (RSU)). With reference to this aspect, device-to-device (D2D) communications are gaining momentum in wireless systems evolution: in 4G standard this concept has been introduced to enhance the coverage via proximity services (ProSe) provisioning or to effectively cope with machine-type communications (MTCs), while in 5G D2D is expected to play a crucial role involving also networking aspects (Malandrino et al. 2014; Shen 2015).

In addition, several use cases have been investigated spanning from emergency, collision, and queue warnings to adaptive cruise control and automated parking. All these services require advanced multicast scheme management, able to jointly support data gathering, fusion, and dissemination within or among (virtual) areas. This involves the clustering of *correlated* vehicle groups according to a hierarchical topology inspired by the *small-world* networking approach and an optimal ad hoc resource allocation policy.

In this entry, we introduce a novel perspective, where a generalized ad hoc paradigm is introduced in cellular networks according to 5G vision to enhance vehicular social networks with the support of both infrastructure and crowd sourcing-based networking while mitigating cost and complexity and increasing scalability,

robustness, and resilience. First, the context-aware vehicular social network paradigm is introduced focusing on their main features and challenges, as well as on opportunities offered by context-aware adaptation. We also discuss how D2D communication schemes could empower social-aware applications in vehicular communications, with specific focus on two use cases (i.e., safety and early warning applications or autonomic transportation systems for *smart cities*). Further, we provide an insight on wireless communication technologies enabling this paradigm shift, with a special regard to beyond 4G systems, mainly LTE-V concept.

Overview on Vehicular Social Networks

Vehicular social networks (VSNs) are an emerging type of social network which allows the exchange of information among drivers, passengers, and vehicles by leveraging dynamic relationships typically built according to vehicles' physical proximity as well as common interests and context (i.e., similar destination) (Yang and Wang 2015).

The joint exploitation of social relationships with emerging communication technologies is considered a promising approach for dramatically improving the added value provided by applications in vehicular networks (e.g., navigation safety applications, navigation efficiency, entertainment, participatory, and urban sensing, emergency (Gerla and Kleinrock 2011)). For instance, ongoing standardization activities on LTE-based V2X communication are specifically focused on supporting the delivery of intelligent services enabled by cooperative awareness (3rd Generation Partnership Project – Technical Specification Group Services and System Aspects 2015). Indeed, considering the social dimension in vehicular networking can bring several advantages. First, since each vehicle can be considered a producer of information, the information exchanged in a VSN may surpass in quality, timing, and local coverage, the one

provided by online remote services. Second, the availability of information on users' common interests and preferences can be exploited to improve the dissemination of information, for instance, by minimizing information delivery to uninterested parties. VSNs can also facilitate the exchange of information, media content, and recommendations for activities related to mobility (e.g., entertainment inside the vehicles for passengers and tourism).

Nonetheless, collaboration in a VSN is not merely limited to application information exchange. Each node (the vehicle itself through the Onboard Unit, OBU, or any other smart device in the vehicle environment which may represent the vehicle or the user digital identity, such as smartphones) conveys sensors, actuators, processing, and communication resources that can contribute to the creation and maintenance of communities targeting the delivery of communication services (opportunistic *networking*) and/or information processing and dissemination services (opportunistic *computing*). Indeed, opportunistic contacts can be exploited to support multi-hop communication schemes (e.g., with or without the assistance of the network infrastructure), thus allowing data exchange between indirect contacts. Moreover, opportunistic networking can ease the cooperation among devices in order to execute processing and information dissemination tasks, by sharing computing, storage, and communication resources for the accomplishment of a given task. On this perspective, VSNs can be progressively configured as a *technological ecosystem* with increasingly autonomous features, which leverage infrastructure resources (e.g., wireless and wired communication infrastructure, remote information services and resources) when needed.

This vision is currently being strengthened by the ongoing evolution of cellular communications toward 5G. Indeed, 5G infrastructure is envisioned as a *neural bearer* where *everything* (applications, processing resources, network, data, etc.) can be provided as a service, and most of intelligence, including context-aware adaptation logic, will be distributed at the edge,

i.e., in the aggregation and access segments up to the end user premises (Soldani and Manzalini 2015). Virtualization technologies are considered key technologies for achieving this objective, with the introduction of cloud platforms at the edge, close to end users, to support the evolution toward cloud-based radio access as well as a novel application delivery approach, targeting faster service delivery and augmented reality (ETSI 2015). Therefore, information processing tasks could be performed locally (on the vehicle and in a proximity-based group of vehicles), at the edge (access/aggregation network), as well as remotely (network operators and service providers' data centers).

As an example, the roadside infrastructure (e.g., the LTE base station site) could host the virtual infrastructure resources for running an application that processes sensor data produced by approaching vehicles, elaborates them, and derives relevant information to be delivered to interested parties with low latency while also delivering data to remote cloud services for resource-demanding processing tasks, such as big data analytics. For instance, a sudden deceleration of a group of vehicles on a road segment could be detected at the roadside infrastructure through collection and real-time analysis of vehicles' speed and location information and trigger the timely delivery of an alert message to incoming vehicles (e.g., possible congestion or obstacle on the road).

In addition to infrastructure-assisted schemes for information processing and dissemination, routing, and communication resource management, also non-infrastructure-assisted schemes can be conceived. Moreover, on top of an opportunistic V2V network, resources hosted by nodes could be exposed as a service and dynamically composed across different nodes for providing added-value application services to the members of the community. For instance, a video captured by a camera on one or more vehicles could be processed by a video transcoder provided by a smartphone, analyzed by an application running on a vehicle OBU, and the results disseminated to the other peers of the network. Figure 1 shows two examples of cooperation at both the appli-

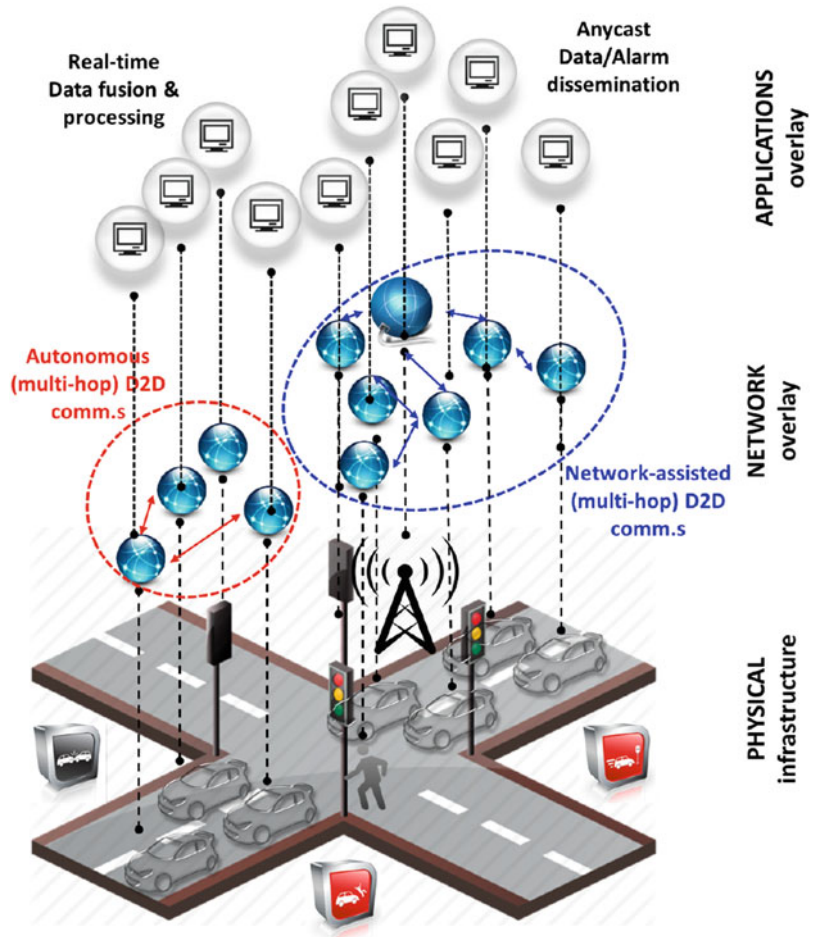
cation and networking layers, enabled by social awareness, as discussed above, where V2X communication is a key enabling technology.

With respect to online and mobile social networks, VSNs pose several specific challenges, mainly due to the fact that the topology of a VSN very often changes and it is formed by peers that move at high speed, then their contact is time limited in the order of tens of seconds, and the technological infrastructure is made by heterogeneous components (Mezghani et al. 2014). In such context, computing trust and guaranteeing trustworthy information sharing in VSNs are challenging requirements (Yang and Wang 2015). Moreover, efficient techniques for handling the dynamic changes in the community structure (i.e., vehicles joining and leaving the community) as well as social-aware routing and communication strategies for the effectively and efficiently disseminating information in a vehicular social network are needed (Maaroufi and Pierre 2014).

Application Scenarios and Communications Use Cases

Cooperation among vehicles can enable a wide range of application services for the benefit of several categories of end users: drivers of private vehicles, drivers of public or collective transport vehicles, transportation authorities, and public administrations, to name a few. Hereafter we provide some scenarios to illustrate how social- and context-aware applications can be empowered by multi-hop D2D communication. To this purpose, we preliminary focus on three different networking use cases, starting from (i) critical information diffusion performed by a single vehicle toward a potential group of neighbors (*anycastransmission*) up to (ii) data exchange among convoys when a communication opportunity arises (*intergroup communication*) and (iii) data aggregation and fusion performed by a vehicle acting as gateway to the Internet (*data gathering and fusion*), as depicted in Fig. 2. As soon as communications become less sporadic and context aware, an ad hoc networking scheme is required: in particular we indicated (i) a basic

Evolution of Vehicular Networks in the Transition from 3G to 4G Systems, Fig. 1 Examples of cooperation at the application and networking layer in vehicular networks



flooding scheme or more advanced solutions relying on (ii) disruption-tolerant networking (DTN) or classical mobile ad hoc networking (MANET).

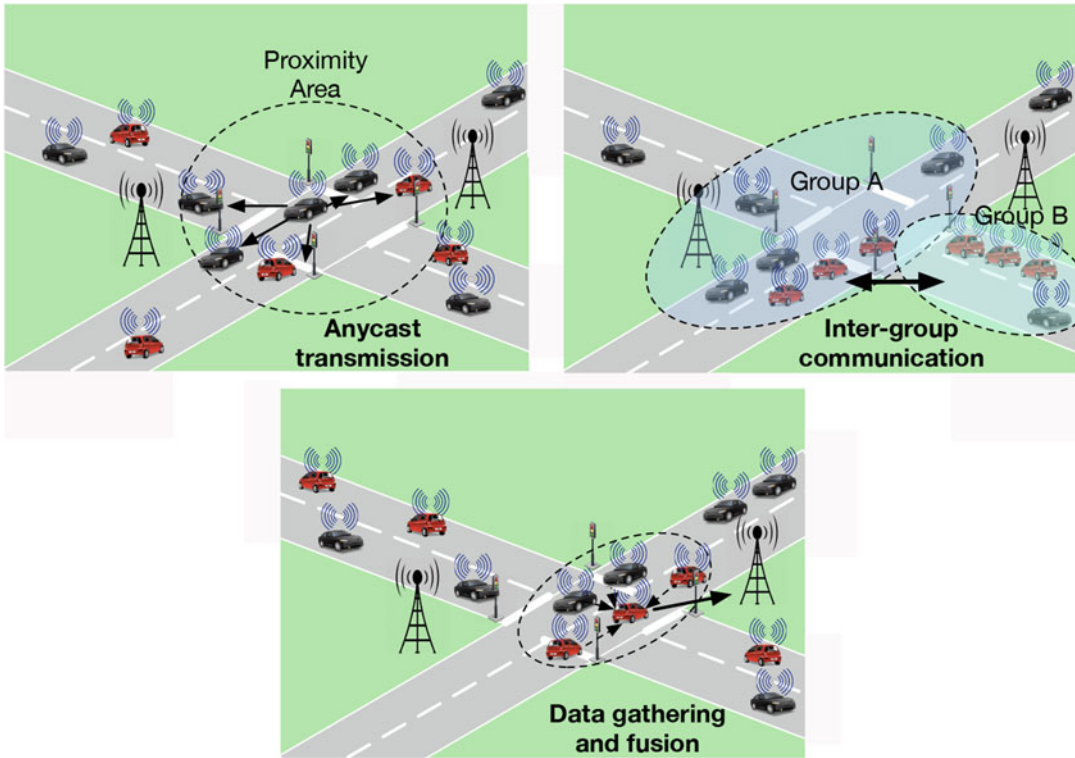
Hereafter we discuss how these networking schemes may support communication requirements of social-aware vehicular applications for safety and early warning and autonomic transport in smart cities.

Safety and Early Warning Applications

Safety and early warning applications are one of the most investigated types of applications in vehicular systems. Basing on the dynamic exchange of context information among OBUs and RSUs, dynamic communities can be built by grouping vehicles that are in the same geographical area and have in common a significant part of

their path or the final destination. For example, events of interest are an abrupt deceleration of (group of) vehicles, an obstacle on the road, a sudden change of weather conditions, and a breakdown of a vehicle. This information can be provided by drivers or passengers (as already enabled by existing applications, e.g., Waze) or detected and automatically *posted* by vehicles so to speed up the time needed for generating and disseminating the alert.

In the data gathering and fusion case, a vehicle, elected as *coordinator* of a group of vehicles, collects context information provided by group members and executes a task for analyzing this information and infers the occurrence of events of interests. For instance, a sudden deceleration and change of direction by a few vehicles may be interpreted as an obstacle on the roadway and



Evolution of Vehicular Networks in the Transition from 3G to 4G Systems, Fig. 2 Potential use cases from a data exchanging and networking perspective

dispatched to other interested peers via an anycast transmission scheme.

Autonomic Transport Systems in Smart Cities

The paradigm of *opportunistic* communication is one of the main ICT enablers in the fields of smart cities. Indeed, adaptive and resilient communications are expected to play a key role in the achievement of smart city goals (i.e., sustainable economic growth, quality of life, rational use of resources through the active involvement and participation of citizens). In particular, V2X communications can enable novel scenarios of cooperation among people, vehicles, and urban infrastructure (e.g., traffic lights, streetlights, message boards, etc.) in order to achieve the best trade-off between private interests (e.g., minimize travel time to work) and public goals (e.g., reduce CO₂ and noise levels). People, vehicles, and smart urban

objects can act as sensors and exchange context information. Distributed or centralized decision approaches can be implemented that based on this information can dynamically adapt the use of urban resources (e.g., change traffic light policies, allow or avoid access to a road). As an example, priority at a crossing road could be dynamically decided according to current situation: e.g., number of vehicles on a lane, a congestion after the crossroad, an approaching ambulance, etc. Drivers can be appropriately alerted through road messages and in vehicle alerts. Furthermore, with the introduction of autonomous vehicle control systems, vehicles could cooperatively agree on a traffic flow policy and autonomously enforce the decision by intergroup communicating (e.g., two groups of vehicles approaching a road intersection on different directions agree on the priority and accordingly decelerate or accelerate).

Enabling Communication Technologies

Wireless communications are still experiencing an impressive evolutionary trend, enabling users to independently collect and share information through *heterogeneous* devices and networks. According to the innovative vision driving the transition to the so-called 5G systems, the simultaneous availability of multi-radio technologies could be advantageously exploited by means of integration and joint management. This increases the networking opportunities leading to a *social* inspired connectivity paradigm. An interesting case study is represented by the so-called vehicular ad hoc networks (VANETs), which aim at revolutionizing the traveling experience especially within a smart city (Rasheed et al. 2013). The underlying concept is indeed to provide each vehicle with a *wireless identity* which is not simply a *global* identifier, as it happens in IPv6, but it is strictly related with user profiles (drivers and passengers) and the context as well. This allows to increase the interoperability and to *extend the horizon* with benefits in terms of safety, quality of experience, and resource optimization. However, this vision strictly depends upon the presence of a communications infrastructure providing connectivity to all the vehicles. In the following we briefly overview the candidate technologies (Karagiannis et al. 2011), spanning from available approaches toward an enhanced paradigm of ad hoc interaction that relies upon next-generation networks.

VANET Architectures

DSRC-WAVE

The available technology for supporting the intelligent transportation system (ITS) applications in the short-range communications is represented by wireless access in vehicular environments (WAVE), also referred to as dedicated short-range communication (DSRC). It is an approved amendment to the IEEE 802.11 standard, whose protocol suite results from the combination of IEEE 802.11p and

IEEE 1609. The former focuses on physical and MAC layers for vehicular environments, while resource management, security, networking, and multichannel operation are handled according to the IEEE 1609 protocol suite.

IEEE 802.11p is based on the popular IEEE 802.11 Wi-Fi, exploiting the same medium access scheme based on the exchange of Request to Send/Clear to Send packets (RTS/CTS) – to alleviate packet collisions in case of unicast communications. However, this technique might not be enabled when a packet is sent in broadcast and the packet collision probability is consequently increased. Furthermore, the quality of service (QoS) support is provided by using the enhanced distributed channel access (EDCA) scheme, which cannot guarantee high packet delivery probability due to the huge number of collisions generated by high priority packets, characterized by a small and fixed contention window (CW) size. The entire 75 MHz bandwidth is divided into seven channels: one channel, i.e., control channel (CCH), is only used to convey emergency and safety traffic, while six service channels (SCH) are occupied by infotainment and multimedia traffic transmission. Further, a channel switching scheme allows the devices to tune to CCH during the CCH time interval and then switch to a specific service channel during the SCH interval.

ISO CALM

Continuous Air interface Long and Medium range (CALM) is the protocol suite standardized by ISO/TC 204 WG for Europe (Mohammad et al. 2011). The CALM CI (communication interface) provides all the functionalities to manage the physical and link layers and can support diverse wireless technologies, ranging from infrared communications to cellular and satellite communications. Moreover, the CALM networking layer coordinates network and transport layers and can either employ IPv6 mobility support protocols or the CALM FAST protocol to perform unicast and broadcast transmissions and store and forward operations.

C2C-CC

Car2Car Communication Consortium (C2C-CC) represents a European open industrial standard mostly focused on (non)safety applications development. It proposed the C2CNet architecture which is not completely IPv6 oriented while supporting multi-hop communications through ad hoc designed routing protocols and vehicles beaconing (Mohammad et al. 2011). Besides, it is inherently conceived for a multi-radio environment supporting both IEEE 802.11a/b/g/n, IEEE 802.11p and cellular technologies.

3G Systems

Presently, the third-generation (3G) networks have been widely deployed and adopted to enable mobile communications. In particular, 3G/UMTS (Universal Mobile Telecommunications System) operates in the band from 1.8 GHz to 2.5 GHz, basically providing data transfer speeds up to 2 Mbps. Besides, 3G HSPA (High-Speed Packet Access) offers a downstream data transfer rate of 14 Mbps and an upstream data transfer rate of 5.74 Mbps. In contrast, 3G HSPA+ (evolved HSPA) achieves 42 Mbps in the downlink and 11 Mbps in the uplink. 3G HSDPA (High-Speed Downlink Packet Access) technique was developed to meet the requirements of bandwidth-intensive applications such as large file transfers and fast Web browsing. It is an ideal technology to support real-time application due to its low latency (70 to 100ms) (Zhao et al. 2013).

To our purpose, it is worth noting that 3G systems provide vehicles with *ubiquitous* Internet access and data communication with other vehicles. These features might be highly beneficial to assist data delivery in VANETs. However, the cost of the 3G traffic is non-negligible, thus limiting their adoption for improving the performance of data delivery in VANETs while limiting the use of this technology to reduce the economical impact. For instance, 3G cellular networks have been commonly used for timely data dissemination to support VANET applications, such as accident prevention and traffic jam avoidance.

Finally, 3G provides an excellent *individual* mobility support, especially in terms of traditional horizontal handoff (OHO) (Lee et al. 2009). However a *seamless* vertical handoff (VHO) extension in order to handle the migration between *heterogeneous* networks is a necessary first step toward VANET, as well as the management of *group* and *constrained* mobility patterns.

4G Standards

The fourth generation (4G)'s most relevant features to VANET scenario are represented by an enhanced radio access technology (RAT) in terms of capacity and latency and a fine-grain spatial coverage via the introduction of *small cells* concept (Feteiha and Hassanein 2015). Nevertheless, the most promising characteristic is the so-called direct communication between devices which is supported under device to device communication (D2D), based on Long Term Evolution (LTE) and LTE-Advanced (LTE-A) (Gupta and Jha 2015; Shen 2015; Agiwal et al. 2016).

D2D paradigm was introduced D2D to reduce the load of evolved node B (eNB) by the 3rd Generation Partnership Project (3GPP) in its Release 12 addressing D2D discovery and communication. The main concept is that devices should be in proximity. D2D is one of the promising applications in terms of network control over communication sessions, where devices have to discover each other and directly communicate with *minimum* involvement of the network. This can solve the problem of delay present in V2V and time required to successfully establish communications as 4G is inherently an all Internet Protocol (IP) technology, which supports packet and circuit switching in its design. Different features were included to support communication systems and networks such as D2D communication which demonstrated its capability to handle V2V under different conditions. Besides, 4G supports high mobility of nodes with speed up to 350 Km/h that can be applied for V2V communication.

However, the spontaneous networking typical of V2V communities requires a dynamic spectrum management, as D2D is an LTE-enabled technology and devices will communicate by

reusing LTE resource. Specifically, the design consideration of D2D is to accommodate vehicles to operate in licensed band without causing harmful interference to licensed users, which can be easily controlled in terms of transmitted power level constraints for a V2V network operating in LTE spectrum (Sakr et al. 2015).

B4G Cellular Networks

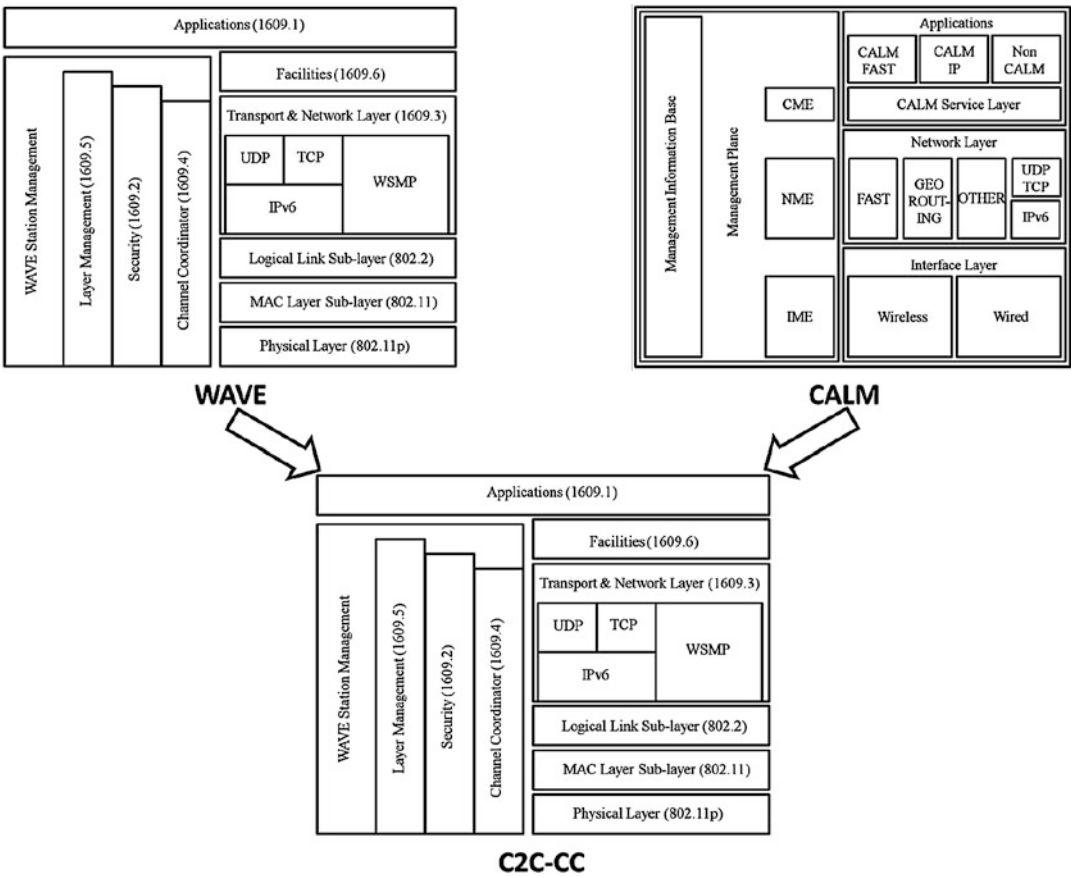
Wide-area advanced 4G systems, especially Long Term Evolution – Advanced (LTE-A) cellular networks, are expected to directly support vehicular applications by means of self-organizing femto-cells, traffic shaping, and *vertical* roaming across different radio access technologies (RATs). In particular LTE networks have extensive coverage for urban and rural scenarios, with low latency and high capacity communications. In addition, LTE allows direct D2D communication which can be easily adopted between vehicles since LTE modems are already integrated into cars. Furthermore, 4G systems intrinsically guarantee long operative life, scalability, cost reduction, and higher performance, despite limitations on the radio access flexibility and on the resource allocation efficiency, especially in the case of sporadic data transmission. As a consequence, cellular systems are being adopted by numerous car manufacturers with the aim of offering diverse services, e.g., remote vehicle monitoring, infotainment, or assisted driving experience. However, though cellular networks are extensively deployed and can easily provide global coverage and ubiquitous connectivity, location-based broadcast applications with stringent requirements in terms of latency and reliability cannot be developed. In addition, a high vehicle density may overload access and core networks and affect the overall QoS of both user and vehicle traffic.

As previously addressed, along with the machine-type communication (MTC) definition, the 3rd Generation Partnership Project (3GPP) Release 12 introduces in LTE-A the support to device-to-device (D2D) or direct mode communications, enabling P2P transmission between devices in proximity (Astely et al. 2013).

This functionality fosters a new generation of devices able to communicate with each other, without any kind of human intervention and without involving the cellular infrastructure for the user plane. On the other hand, D2D communications pose also several challenges. First of all, direct mode communications should be established without affecting traditional communications via base station, i.e., by avoiding possible *interferences* with other devices. Secondly, peer and service discovery functionality need to be introduced, since a device is typically not aware of other terminals in the proximity. Following the paradigm of multi-hop cellular networks (MCN) (Nishiyama et al. 2014), D2D also provides more flexibility in a cellular network, as direct links enable the communication among the terminals using multiple hops.

5G Systems

Mobile network evolution is expected to converge by 2020 into the novel 5G system effectively integrating both wired and wireless segments to accomplish expanded use cases and applications (Wang et al. 2014; Boccardi et al. 2014). In particular, new services are expected such as the massive sensor and vehicular-to-everything communications, according to the Internet of Things (IoT) visions, which refers to interconnection and the autonomous information exchange among a *community* of generic devices. The latter aspect is involved with the so-called safety critical applications, requiring extremely low setup times and transmission delays. 5G is conceived as an unplanned *ultradense* system, where nodes are able to self-organize and to improve cell performance. This trend is clearly pointed out by the emphasis 3GPP working group is progressively posing on *small* cells, whose limit is represented by the D2D technology, providing an optimal coordination/cooperation among devices/cells. As the system is assumed to be inherently heterogeneous, a higher degree of flexibility is required for both radio access and backhaul networks. 5G systems are expected to offer native support for automotive-related



Evolution of Vehicular Networks in the Transition from 3G to 4G Systems, Fig. 3 VANET candidate protocol stack comparisons (Mohammad et al. 2011)

communications, provided that the LTE-V concept will be defined in Release 13 and 14 by taking into account the LTE-D aspects which facilitate the devices and services discovery and the infrastructure fault handling in a partial stand-alone network arrangement. To this purpose, VANET represents a relevant case study for 5G architecture, as it is highlighted by comparatively presenting WAVE, CALM, and C2C-CC functional stacks (Fig. 3).

Conclusions

In this entry, we provided an outlook on the potential impact of opportunistic ad hoc communication in VANET technologies

and applications, with a specific focus on VSNs. First, we analyzed the relevance of social aspects in vehicular applications and networking and the role of context awareness in VSNs in triggering appropriate adaptation actions at the application as well as at the communication level. Then, we reviewed candidate technologies both in the ad hoc and cellular domains that enable this paradigm shift with a special focus on the transition between 3G and 4G, embracing 5G standard evolution. We also characterized the D2D paradigm pointing out the relevant use cases where the benefits provided by multi-hop D2D communication schemes to emerging VSN applications could be provided in the near future.

References

- 3rd Generation Partnership Project – Technical Specification Group Services and System Aspects (2015) 3GPP TR 22.885 – Study on service aspects for LTE-based V2X. Technical report, 3rd generation partnership project
- Agiwal M, Roy A, Saxena N (2016) Next generation 5g wireless networks: a comprehensive survey. *IEEE Commun Surv Tutor* 18(3):1617–1655
- Astely D, Dahlman E, Fodor G, Parkvall S, Sachs J (2013) LTE release 12 and beyond. *Commun Mag IEEE* 51(7):154–160
- Boccardi F, Heath R, Lozano A, Marzetta T, Popovski P (2014) Five disruptive technology directions for 5g. *Commun Mag IEEE* 52(2):74–80
- ETSI (2015) Mobile edge computing (MEC); service scenarios – ETSI GS MEC-IEG 004 V1.1.1. Technical report., ETSI
- Feteiha MF, Hassanein HS (2015) Enabling cooperative relaying vanet clouds over LTE-a networks. *IEEE Trans Veh Technol* 64(4):1468–1479
- Gerla M, Kleinrock L (2011) Vehicular networks and the future of the mobile internet. *Comput Netw* 55(2):457–469
- Gupta A, Jha RK (2015) A survey of 5g network: architecture and emerging technologies. *IEEE Access* 3:1206–1232
- Karagiannis G, Altintas O, Ekici E, Heijenk G, Jarupan B, Lin K, Weil T (2011) Vehicular networking: a survey and tutorial on requirements, architectures, challenges, standards and solutions. *Commun Surv Tutor* 13(4):584–616
- Lee S, Sriram K, Kim K, Kim YH, Golmie N (2009) Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks. *IEEE Trans Veh Technol* 58(2):865–881
- Maaroufi S, Pierre S (2014) Vehicular social systems: an overview and a performance case study. In: *Proceedings of the fourth ACM international symposium on development and analysis of intelligent vehicular networks and applications, DIVANet '14*, pp 17–24
- Malandrino F, Casetti C, Chiasserini CF (2014) Toward d2d-enhanced heterogeneous networks. *Commun Mag IEEE* 52(11):94–100
- Mezghani F, Dhaou R, Nogueira M, Beylot AL (2014) Content dissemination in vehicular social networks: taxonomy and user satisfaction. *Commun Mag IEEE* 52(12):34–40
- Mohammad S, Rasheed A, Qayyum A (2011) Vanet architectures and protocol stacks: a survey. In: Strang T, Festag A, Vinel A, Mehmood R, Rico Garcia C, Rckl M (eds) *Communication technologies for vehicles*, lecture notes in computer science, vol 6596. Springer, Berlin/Heidelberg, pp 95–105. https://doi.org/10.1007/978-3-642-19786-4_9
- Nishiyama H, Ito M, Kato N (2014) Relay-by-smartphone: realizing multihop device-to-device communications. *Commun Mag IEEE* 52(4):56–65
- Rasheed A, Zia H, Hashmi F, Hadi U, Naim W, Ajmal S (2013) Fleet and convoy management using VANET. *J Comput Netw* 1:1–9
- Sakr AH, Tabassum H, Hossain E, Kim DI (2015) Cognitive spectrum access in device-to-device-enabled cellular networks. *IEEE Commun Mag* 53(7):126–133
- Shen X (2015) Device-to-device communication in 5g cellular networks. *Netw IEEE* 29(2):2–3
- Soldani D, Manzalini A (2015) Horizon 2020 and beyond: on the 5g operating system for a true digital society. *Veh Technol Mag IEEE* 10(1):32–42
- Wang CX, Haider F, Gao X, You XH, Yang Y, Yuan D, Aggoune H, Haas H, Fletcher S, Hepsaydir E (2014) Cellular architecture and key technologies for 5g wireless communication networks. *Commun Mag IEEE* 52(2):122–130
- Yang Q, Wang H (2015) Toward trustworthy vehicular social networks. *Commun Mag IEEE* 53(8):42–47
- Zhao Q, Zhu Y, Chen C, Zhu H, Li B (2013) When 3g meets vanet: 3g-assisted data delivery in vanets. *IEEE Sensors J* 13(10):3575–3584

Expected Utility Theory

► Prospect Theory for Communication Networks

Exposed Node Problem

► Hidden and Exposed Terminal Problems

Extending WSN Lifetime Based on Evolutionary Clustering Algorithm

Alaa Sheta¹ and Basma Solaiman²

¹Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

²New and Renewable Energy Authority, Cairo, Egypt

Synonyms

Computational intelligence; Energy consumption; Energy saving; Machine learning; Soft-computing

Definition

The wireless sensor network (WSN) is a network of tens or hundreds of computing nodes (i.e., sensors) with various processing capabilities distributed over a large landscape. The communication inside the network is formed using wireless transmission techniques. The nodes have the capability to measure and process external environmental data and then send the data messages to a base station (BS) for further handling.

Historical Background

In the mid-1950s, the US Army developed the first WSN, called the “Sound Surveillance System (SOSUS),” for tracking and surveillance of Soviet submarines in the Atlantic and Pacific oceans. The earliest real-world project founded, in the 1970s, was the Automated Local Evaluation in Real Time (ALERT). It was designed to detect the existence of flood using sensors that take measurements as: temperature, humidity, rain, and water level (Golen et al. 2009). Recently, WSN is applied to civil applications, medical health care (Schwiebert et al. 2001), smart energy (Rupinder Kaur 2012), and industrial automation (Marketsandmarkets 2014). The global industrial wireless sensor network market size is estimated to grow from \$401.23 Million in 2013 to \$944.92 million by 2020. Figure 1 shows two layouts, a WSN military application proposed in Razaque and Elleithy (2014) and another layout for WSN health-care monitoring system presented in Saleem et al. (2011).

Foundation

To design a reliable WSN, many challenging parameters facing the design have to be considered as routing, deployment, and localization. WSN lifetime is considered one of the main challenges facing WSN design. The primary limiting factor for the sensor network is the energy supply which directly affects the WSN’s lifetime. WSN sensors are normally equipped with a lim-

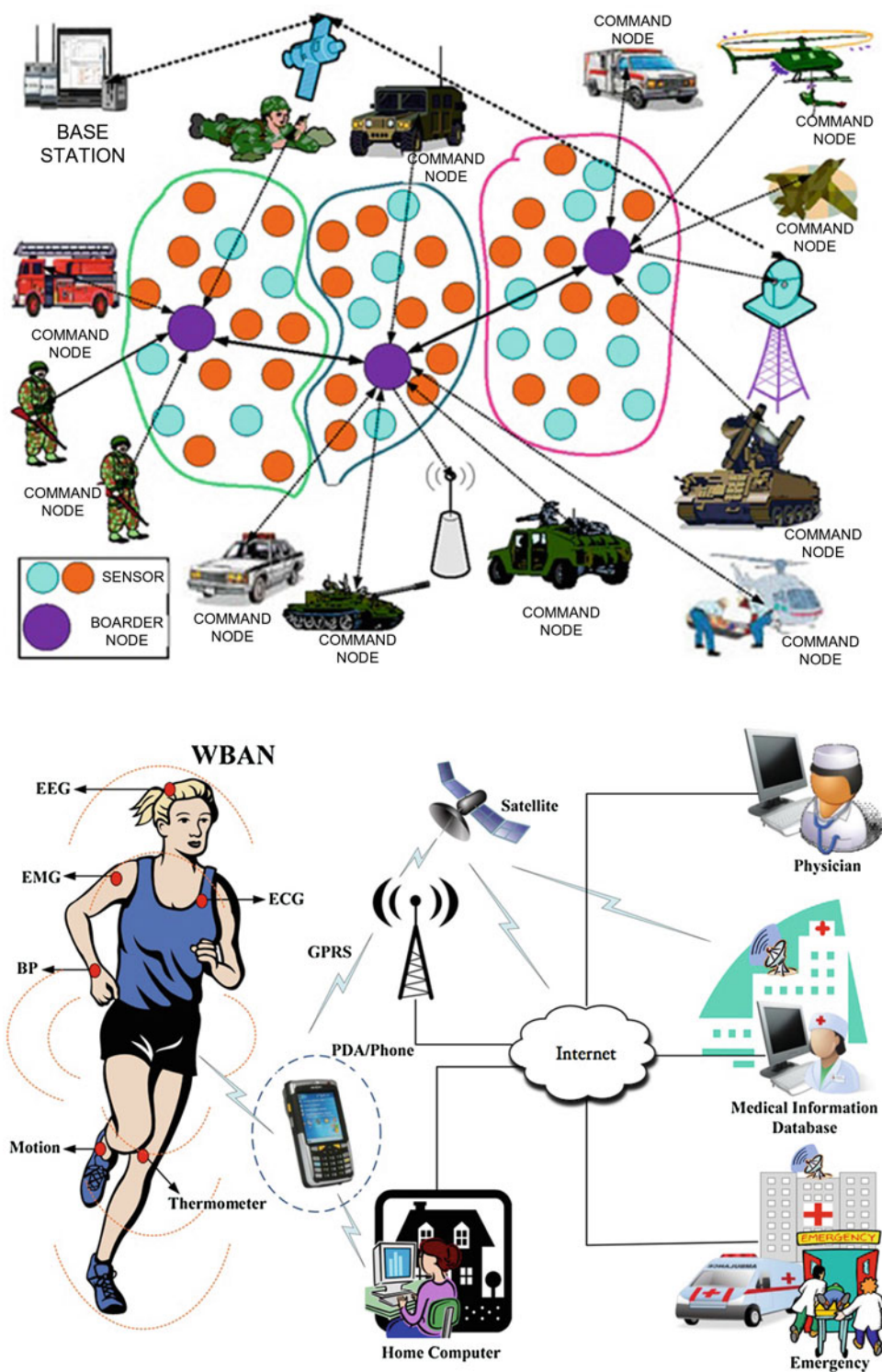
ited capacity battery. Since the nodes could be deployed in hostile areas, it may be impossible to replace or recharge the nodes’ batteries. Thus, the design of the network must consider methods for maximizing the network lifetime and minimizing the energy consumption.

WSN clustering was proposed as a solution for the energy problem (Li et al. 2005; Vinay Kumar and Tiwari 2011; Liao and Zhu 2013). Clustering allows the network to work in a hierarchical manner. Therefore, WSN can incorporate efficient utilization of limited resources of the sensor nodes and extend its lifetime. Clustering process divides the network into disjoint groups called clusters. Each cluster elects only one node to become a cluster head (CH). Each CH is responsible for collecting sensed data from all nodes in the cluster group, concatenating it into a single message, and sending the data message to the BS (Khanna et al. 2006; Abbasi and Younis 2007). Clustering serves in reducing communication overhead (Patole 2012) and scalability and providing better use of network resources. The adoption of clustering as a technique for energy consumption and extending lifetime raises many design questions such as:

1. How many clusters should we have?
2. How many CH is adequate for a specific layout?
3. Which node should be elected as a CH?
4. Which cluster a node should join?

To answer the above-described questions, a powerful search algorithm needs to be used to explore the space of all possible design parameters described by these four questions. Traditional techniques, as the famous low-energy adaptive clustering hierarchy (LEACH) protocol (Heinzelman 2000), aimed to provide some answers of these questions in a random manner. However, they are expensive algorithms and suffer from nonuniformity in clusters and CH distribution (Hou et al. 2010).

Optimization algorithms were provided to solve various problems associated to WSN design and deployment (Madan et al. 2007). Among these algorithms are the Evolutionary



Extending WSN Lifetime Based on Evolutionary Clustering Algorithm, Fig. 1 Top image: WSN military application (Razaque and Elleithy 2014). Bottom image: WSN health-care monitoring (Saleem et al. 2011)

Algorithms (EAs) (Carballido et al. 2007; Mehr 2011; Sheta and Solaiman 2015). EA are algorithms with stochastic components that aim for finding optimal parameters for WSN design such as the adequate number of cluster heads (Solaiman and Sheta 2015b). EA are characterized by two major components: exploration and exploitation. Exploration enables the algorithm to diverse in the search space globally till it finds promising solution space. Exploitation refers to intensely search the promising region for a good solution. Putting great potential in exploitation leads to getting trapped in a local optimum solution, while using exploration reduces the chances of population convergence and causes deviation from the best solution. Therefore, EA has to maintain a good balance between exploration and exploitation to succeed in reaching a high-quality solution. EA performance is robust irrespective of the dimension of the search space (Golen et al. 2009; Martins et al. 2010). They can be implemented easily using any computer language and require minimal parameter tuning effort. However, there is no guarantee that optimal solutions can be reached. Genetic algorithm (GA), particle swarm optimization, (PSO), and differential evolution (DE) are well-known EA algorithms that have solved many optimization problems efficiently (Siew et al. 2012; Singh and Lobiyal 2012). A summary of these three algorithms is provided.

- Genetic algorithm was presented by J. Holland (1975) and extensively studied by Goldberg (1989; 2002), DeJong (1975; 1980; 1988), and others. GA successfully handled many areas of applications and was able to solve a wide variety of difficult numerical optimization problems. GA is inspired from Charles Darwin's theory of evolution: "the survival of the fittest" where only the best individuals are able to survive in harsh biological conditions. Thus, as generations run, better individuals are obtained by reproduction genetic operations as crossover and mutation of chromosomes.
- Particle swarm optimization was developed in 1995 (Kennedy and Eberhart 1995) by James Kennedy (social psychologist) and Russell Eberhart (electrical engineer). PSO is a robust stochastic nonlinear optimization technique based on movement and intelligence of swarms. It is inspired from social behavior of flock of birds or fish. The flock randomly search for food in an area by interacting together and following the nearest bird to the food. PSO consists of number of individuals, called particles. Each particle represents a possible solution. The particles move in the search space looking for the optimal solution. The particles update their position according the topology that allows each particle to exchange information with its neighborhood particles to memorize the best position reached by the swarm.
- Differential evolution is a very simple evolutionary algorithm that has been developed by Price and Storn in 1995 (Storn and Price 1997). DE is accurate and converges fast to the domain of optimal solution. It replaces classical crossover and mutation operators in GA with alternative operators. It attempts to replace crossover operator of GA by a special type of differential operator for reproducing offspring in the next generation. The population of DE is represented by its position in the search space, named vectors (Das et al. 2008). While GA relies on crossover, DE differs from GA in the order of operators used and uses mutation as the primary search mechanism.

Table 1 shows number of research work in the area of WSN clustering using EA. More details on WSN components and clustering algorithms are given in Solaiman and Sheta (2013). Hybrid EA clustering algorithms resulted in decreased computational effort and showed more adaptability to the dynamic nature of WSN (Solaiman and Sheta 2015,b).

Figure 2 shows part of the developed results in (Solaiman and Sheta 2015b; Sheta and Solaiman

Extending WSN Lifetime Based on Evolutionary Clustering Algorithm, Table 1 A list of previous works from year 2001 to 2016 that have applied evolutionary search algorithms for WSN clustering

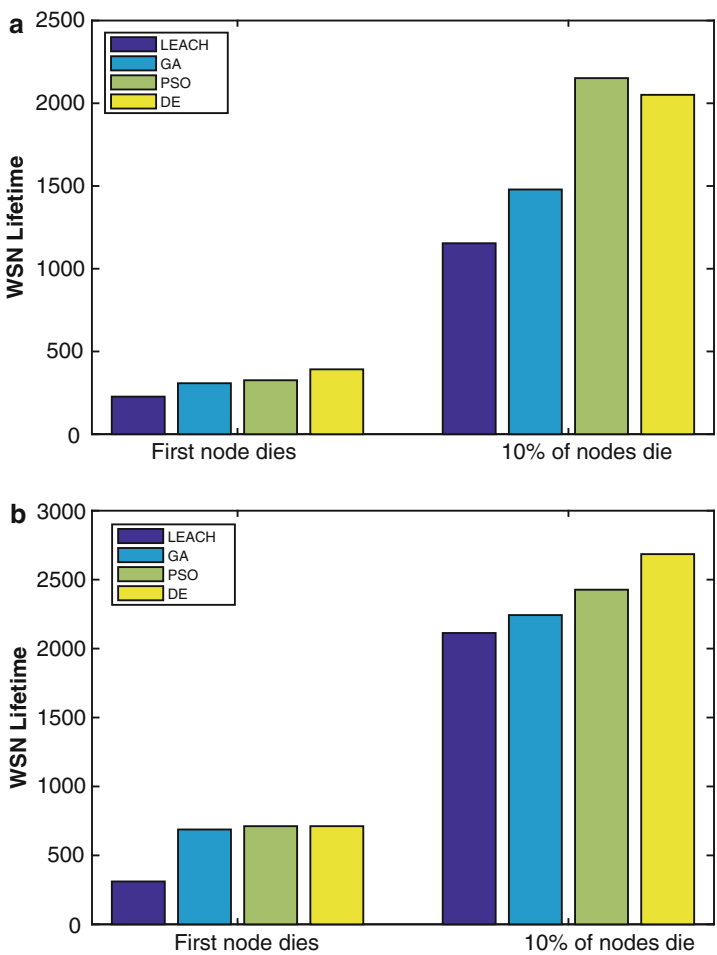
#	Contribution	Year	Model	Aspects
1	Distributed sensor placement with sequential particle swarm optimization (Ngatchou et al. 2005)	2005	PSO	Deployment
2	Optimized sink node path using particle swarm optimization (Mendis et al. 2006)	2006	PSO	Deployment
3	A study of particle swarm optimization in urban traffic surveillance system (Hu et al. 2006)	2006	PSO	Deployment
4	CGD-GA: a graph-based genetic algorithm for sensor network design (Carballido et al. 2007)	2007	GAs	Deployment
5	Differential evolution based deployment of wireless sensor networks (Ayinde and Barnawi 2014)	2014	DE	Deployment
6	Adaptive routing in wireless communication networks using swarm intelligence (Arabshahi et al. 2001)	2001	PSO	Routing
7	A parallel ant colony optimization algorithm for all-pair routing in MANETs (Islam et al. 2003)	2003	ACO	Routing
8	Multi-objective routing in wireless sensor networks with a differential evolution algorithm (Xue et al. 2006)	2006	PSO	Routing
9	Genetic algorithm for energy efficient clusters in wireless sensor networks (Hussain et al. 2007b)	2007	GAs	Routing
10	Estimation of node localization with a real-coded genetic algorithm in WSNs (Nan et al. 2007)	2007	GAs	Localization
11	Two-phase stochastic optimization to sensor network localization (Marks and Niewiadomska-Szynkiewicz 2007)	2007	GAs	Localization
12	Localization in wireless sensor networks using particle swarm optimization (Gopakumar and Jacob 2008)	2008	PSO	Localization
13	Bio-inspired node localization in wireless sensor networks (Kulkarni et al. 2009)	2009	PSO	Localization
14	Differential evolution approach for localization in wireless sensor networks (Harikrishnan et al. 2014)	2014	DE	Localization
15	Cluster-head identification in ad hoc sensor networks using particle swarm optimization (Tillett et al. 2002)	2002	PSO	Clustering
16	Sensor network optimization using a genetic algorithm (Jin et al. 2003)	2003	GAs	Clustering
17	Genetic algorithm for energy efficient clusters in wireless sensor networks (Hussain et al. 2007)	2007	GAs	Clustering
18	Energy-aware clustering for wireless sensor networks using particle swarm optimization (Latiff et al. 2007)	2007	PSO	Clustering
19	Optimizing the communication distance of an ad hoc wireless sensor networks by genetic algorithms (Al-Obaidy et al. 2008)	2008	GAs	Clustering
20	Evolutionary genetic algorithm for efficient clustering of wireless sensor networks (Seo et al. 2009)	2009	GAs	Clustering
21	Intelligent clustering in wireless sensor networks (Heidari and Movaghar 2009)	2009	GAs	Clustering
22	Clustering strategy of wireless sensor networks based on improved discrete particle swarm optimization (Hou et al. 2010)	2010	PSO	Clustering
23	Double cluster-heads clustering algorithm for wireless sensor networks using PSO (Ruihua et al. 2011)	2011	PSO	Clustering
24	Particle swarm optimization based clustering in wireless sensor networks: The effectiveness of distance altering (Bennani and El Ghanami 2012)	2012	PSO	Clustering
25	Multiple mobile agents in wireless sensor networks using genetic algorithms (LingaRaj.K 2012)	2012	GAs	Clustering

(continued)

Extending WSN Lifetime Based on Evolutionary Clustering Algorithm, Table 1 (Continued)

26	Optimizing cluster-head selection in wireless sensor networks using genetic algorithm and harmony search algorithm (Karimi et al. 2012)	2012	GAs	Clustering
27	Energy conscious clustering of wireless sensor network using PSO incorporated cuckoo search (Karthikeyan and Venkatalakshmi 2012)	2012	PSO, CS	Clustering
28	Hierarchical clustering using genetic algorithm in wireless sensor networks (Shurman et al. 2013)	2013	GAs	Clustering
29	Energy aware optimal cluster head selection in wireless sensor networks (Natarajan et al. 2013)	2013	PSO	Clustering
30	A novel differential evolution based clustering algorithm for wireless sensor networks (Kuila and Jana 2014)	2014	DE	Clustering
31	A novel clustering algorithm based on particle swarm optimization for wireless sensor networks (Jing et al. 2014)	2014	PSO	Clustering
32	Evolving Clustering Algorithms for wireless sensor networks with various radiation patterns to reduce energy consumption (Solaiman and Sheta 2016)	2016	PSO	Clustering

Extending WSN Lifetime Based on Evolutionary Clustering Algorithm, Fig. 2 WSN lifetime for (a) 100 nodes and (b) 50 nodes with 2J initial energy



2015) describing the lifetime for WSN with 100 and 50 randomly deployed sensor nodes. The graphs show the EA's results in optimizing the WSN parameters (i.e., the number of sensors in each cluster, the selection of the number of CH, and the elected CHs) for a WSN with extended

lifetime compared to the LEACH protocol. EAs proved to a better layout and prolonging the WSN lifetime.

Key Applications

Extending WSN lifetime is major challenge that has to be considered in designing the network since the WSN applications are remotely deployed and battery replacement is impossible.

Cross-References

- [Application of Machine Learning in Wireless Sensor Network](#)
- [Machine Learning in Wireless Sensor Networks for the Internet of Things](#)
- [Machine Learning Paradigms in Wireless Network Association](#)

References

- Abbasi AA, Younis M (2007) A survey on clustering algorithms for wireless sensor networks. *Comput Commun* 30(14–15):2826–2841
- Al-Obaidy M, Ayesh A, Sheta AF (2008) Optimizing the communication distance of an ad hoc wireless sensor networks by genetic algorithms. *Artif Intell Rev* 29(3):183–194
- Arabshahi P, Gray A, Kassabalidis I, El-Sharkawi MA, Marks RJ, Das A, Narayanan S (2001) Adaptive routing in wireless communication networks using swarm intelligence. In: *International communications satellite systems conference*, Toulouse
- Ayinde BO, Barnawi AY (2014) Differential evolution based deployment of wireless sensor networks. In: *2014 IEEE/ACS 11th international conference on computer systems and applications (AICCSA)*, Doha, pp 131–137. <https://doi.org/10.1109/AICCSA.2014.7073189>
- Bennani K, El Ghanami D (2012) Particle swarm optimization based clustering in wireless sensor networks: the effectiveness of distance altering. In: *Proceedings of the international conference on complex systems*, Agadir, pp 1–4
- Carballido JA, Ponzoni I, Brignole NB (2007) CGD-GA: a graph-based genetic algorithm for sensor network design. *Inf Sci* 177(22):5091–5102. <https://doi.org/10.1016/j.ins.2007.05.036>
- Das S, Abraham A, Konar A (2008) Particle swarm optimization and differential evolution algorithms: technical analysis, applications and hybridization perspectives. In: Liu Y, Sun A, Loh H, Lu W, Lim EP (eds) *Advances of computational intelligence in industrial systems. Studies in computational intelligence*, vol 116. Springer, Berlin/Heidelberg, pp 1–38. http://doi.org/10.1007/978-3-540-78297-1_1
- De Jong K (1988) Learning with genetic algorithms: an overview. *Mach Learn* 3(3):121–138
- De Jong KA (1975) Analysis of behavior of a class of genetic adaptive systems. PhD thesis, University of Michigan, Ann Arbor
- De Jong KA (1980) Adaptive system design: a genetic approach. *IEEE Trans Syst Man Cybern* 10(3):556–574
- Goldberg D (2002) *The design of innovation: lessons from and for competent genetic algorithms*. Kluwer Academic, Boston
- Goldberg DE (1989) *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Professional, Reading
- Golen EF, Yuan B, Shenoy N (2009) Underwater sensor deployment using an evolutionary algorithm. In: *Proceedings of the 2009 international conference on wireless communications and mobile computing: connecting the world wirelessly*. Leipzig, pp 1141–1145. <https://doi.org/10.1145/1582379.1582630>
- Gopakumar A, Jacob L (2008) Localization in wireless sensor networks using particle swarm optimization. In: *2008 IET international conference on wireless, mobile and multimedia networks*, Mumbai, pp 227–230. <https://doi.org/10.1049/cp:20080185>
- Harikrishnan R, Kumar VJS, Ponmalar PS (2014) Differential evolution approach for localization in wireless sensor networks. In: *2014 IEEE international conference on computational intelligence and computing research*, Coimbatore, pp 1–4. <https://doi.org/10.1109/ICCIC.2014.7238536>
- Heidari E, Movaghar A (2009) Intelligent clustering in wireless sensor networks. In: *First international conference on networks and communications*, pp 12–17
- Heinzelman WB (2000) *Application-specific protocol architectures for wireless networks*. PhD thesis, Massachusetts Institute of Technology
- Holland JH (1975) *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor
- Hou J, Fan X, Wang W, Jie J, Wang Y (2010) Clustering strategy of wireless sensor networks based on improved discrete particle swarm optimization. In: *Proceedings of the 2010 sixth international conference on natural computation (ICNC)*, vol 7, Yantai, pp 3866–3870. <https://doi.org/10.1109/ICNC.2010.5582664>
- Hu J, Song J, Kang X, Zhang M (2006) A study of particle swarm optimization in Urban traffic surveillance system. In: *IMACS multicongress computational engineering system Application*, Beijing
- Hussain S, Matin A, Islam O (2007) Genetic algorithm for energy efficient clusters in wireless sensor networks. In: *Proceedings of the fourth international conference on information technology*, Las Vegas, NV, pp 147–154

- Hussain S, Matin AW, Islam O (2007b) Genetic Algorithm for Hierarchical Wireless Sensor Networks. *Journal of Networks* 2(5):87–97
- Islam MT, Thulasiraman P, Thulasiram RK (2003) A parallel ant colony optimization algorithm for all-pair routing in MANETs. In: *Proceedings international parallel and distributed processing symposium, Nice*, 8pp. <https://doi.org/10.1109/IPDPS.2003.1213470>
- Jin S, Zhou M, Wu AS (2003) Sensor network optimization using a genetic algorithm. In: *Proceedings of the 7th world multiconference on systemics, cybernetics, and informatics*. Orlando, FLorida
- Jing Z, Le T, Shuaibing Z (2014) A novel clustering algorithm based on particle swarm optimization for wireless sensor networks. In: *Proceedings of the 26th Chinese control and decision conference*, Changsha, pp 2769–2772
- Karimi M, Naji H, Golestani S (2012) Optimizing cluster-head selection in wireless sensor networks using genetic algorithm and harmony search algorithm. In: *Proceedings of the 20th Iranian conference on electrical engineering*. Tehran, pp 706–710
- Karthikeyan M, Venkatalakshmi K (2012) Energy conscious clustering of wireless sensor network using PSO incorporated cuckoo search. In: *Proceedings of the third international conference on computing communication networking technologies*, Coimbatore, pp 1–7
- Kennedy J, Eberhart R (1995) Particle swarm optimization. In: *Proceedings of the IEEE international conference on neural networks*, vol 4, Perth, WA, pp 1942–1948
- Khanna R, Liu H, Chen HH (2006) Self-organization of sensor networks using genetic algorithms. In: *Proceedings of the IEEE international conference on communications*, vol 8, Istanbul, pp 3377–3382
- Kuila P, Jana PK (2014) A novel differential evolution based clustering algorithm for wireless sensor networks. *Appl Soft Comput* 25:414–425. <https://doi.org/10.1016/j.asoc.2014.08.064>, <http://www.sciencedirect.com/science/article/pii/S156849461400430X>
- Kulkarni RV, Venayagamoorthy GK, Cheng MX (2009) Bio-inspired node localization in wireless sensor networks. In: *2009 IEEE international conference on systems, man and cybernetics*, San Antonio, TX, pp 205–210. <https://doi.org/10.1109/ICSMC.2009.5346107>
- Latiff N, Tsimenidis C, Sharif B (2007) Energy-aware clustering for wireless sensor networks using particle swarm optimization. In: *Proceedings of the IEEE 18th international symposium on personal, indoor and mobile radio communications*, Athens, pp 1–5
- Li C, Ye M, Chen G, Wu J (2005) An energy-efficient unequal clustering mechanism for wireless sensor networks. In: *Proceedings of IEEE international conference on mobile adhoc and sensor systems*, Washington, DC, pp 1–8
- Liao Q, Zhu H (2013) An energy balanced clustering algorithm based on leach protocol. In: *Proceedings of the 2nd international conference on systems engineering and modeling*, IEEE, advances in intelligent systems research. Beijing, <https://doi.org/10.2991/icsem.2013.15>
- LingaRaj KN, Aradhana D (2012) Multiple mobile agents in wireless sensor networks using genetic algorithms. *Int J Sci Eng Res* 3(8):1–5
- Madan R, Cui S, Lall S, Goldsmith AJ (2007) Modeling and optimization of transmission schemes in energy-constrained wireless sensor networks. *IEEE/ACM Trans Netw* 15(6):1359–1372. <https://doi.org/10.1109/TNET.2007.897945>
- Marketsandmarkets (2014) IWSN (Industrial Wireless Sensor Network) market by sensor, technology, application, geography trend, and forecast to 2020. http://www.marketsandmarkets.com/Market-Reports/wireless-sensor-networks-market-445.html?clid=CNSm3ui45cMCFWfJtAodxGkA_g
- Marks M, Niewiadomska-Szynkiewicz E (2007) Two-phase stochastic optimization to sensor network localization. In: *2007 international conference on sensor technologies and applications (SENSORCOMM 2007)*, Valencia, pp 134–139
- Martins FVC, Carrano EG, Wanner EF, Takahashi RHC, Mateus GR (2010) An evolutionary dynamic approach for designing wireless sensor networks for real time monitoring. In: *Proceedings of the 2010 IEEE/ACM 14th international symposium on distributed simulation and real time applications*. IEEE Computer Society, Washington, pp 161–168. <https://doi.org/10.1109/DS-RT.2010.25>
- Mehr MA (2011) Design and implementation a new energy efficient clustering algorithm using genetic algorithm for wireless sensor networks. *World Acad Sci Eng Technol* 52:430–433
- Mendis C, Guru S, Halgamuge S, Fernando S (2006) Optimized sink node path using particle swarm optimization. In: *20th international conference of advanced information networking applications AINA*, Vienna
- Nan GF, Li MQ, Li J (2007) Estimation of node localization with a real-coded genetic algorithm in WSNs. In: *2007 international conference on machine learning and cybernetics*, vol 2, Hong Kong, pp 873–878. <https://doi.org/10.1109/ICMLC.2007.4370265>
- Natarajan M, Arthi R, Murugan K (2013) Energy aware optimal cluster head selection in wireless sensor networks. In: *Proceedings of the fourth international conference on computing, communications and networking technologies*, pp 1–4
- Ngatchou PN, Fox WLJ, El-Sharkawi MA (2005) Distributed sensor placement with sequential particle swarm optimization. In: *Proceedings 2005 IEEE swarm intelligence symposium, SIS'2005*, pp 385–388
- Patole JR (2012) Clustering in wireless sensor network using K-MEANS and MAP REDUCE algorithm. Master's thesis, Department of Computer Engineering and Information Technology, College of Engineering, Pune
- Razaque A, Elleithy KM (2014) Energy-efficient border node medium access control protocol for wireless sensor networks. *Sensors* 14:5074–5117

- Ruihua Z, Zhiping J, Xin L, Dongxue H (2011) Double cluster-heads clustering algorithm for wireless sensor networks using PSO. In: Proceedings of the 6th IEEE conference on industrial electronics and applications, Beijing, pp 763–766
- Rupinder Kaur I MS (2012) Efficient energy consumption in wireless sensor networks using optimization technique. *Int J Eng Sci Technol* 2(5): 1290–1294
- Saleem S, Ullah S, Kwak KS (2011) A study of IEEE 802.15.4 security framework for wireless body area network. *CoRR* abs/1102.0682. <http://arxiv.org/abs/1102.0682>, 1102.0682
- Schwiebert L, Gupta SK, Weinmann J (2001) Research challenges in wireless networks of biomedical sensors. In: Proceedings of the 7th annual international conference on mobile computing and networking, MobiCom01. ACM, pp 151–165. <https://doi.org/10.1145/381677.381692>
- Seo HS, Oh SJ, Lee CW (2009) Evolutionary genetic algorithm for efficient clustering of wireless sensor networks. In: Proceedings of the 6th IEEE consumer communications and networking conference, Las Vegas, NV, pp 1–5
- Sheta AF, Solaiman B (2015) Evolving clustering algorithms for wireless sensor networks with various radiation patterns to reduce energy consumption. In: Science and information conference (SAI), London, pp 1037–1045. <https://doi.org/10.1109/SAI.2015.7237270>
- Shurman M, Al-Mistarihi M, Mohammad A, Darabkh K, Ababnah A (2013) Hierarchical clustering using genetic algorithm in wireless sensor networks. In: Proceedings of the 36th international convention on information communication technology electronics microelectronics, Opatija, pp 479–483
- Siew ZW, Wong CH, Chin CS, Kiring A, Teo K (2012) Cluster heads distribution of wireless sensor networks via adaptive particle swarm optimization. In: Proceedings of the fourth international conference on computational intelligence, communication systems and networks, Phuket, pp 78–83
- Singh B, Lobiyal D (2012) A novel energy-aware cluster head selection based on particle swarm optimization for wireless sensor networks. *Human-centric Comput Inf Sci* 2(1):13. <https://doi.org/10.1186/2192-1962-2-13>
- Solaiman B, Sheta A (2013) Computational intelligence for wireless sensor networks: applications and clustering algorithms. *Int J Comput Appl* 73(15):1–8
- Solaiman B, Sheta A (2015) Energy optimization in wireless sensor networks using a hybrid K-Means PSO clustering algorithm. *Turk J Electr Eng Comput Sci* 24:2679–2695
- Solaiman B, Sheta A (2015b) Evolving a hybrid k-means clustering algorithm for wireless sensor network using PSO and GA. *Int J Comput Sci Issues* 12(1):23–32
- Solaiman B, Sheta A (2016) Evolving a clustering algorithm for wireless sensor network using particle swarm optimisation. *Int J Swarm Intell* 2(1):43–65
- Storn R, Price K (1997) Differential evolution – a simple and efficient adaptive scheme for global optimization over continuous spaces. *J Glob Optim* 11(4):341–359
- Tillett J, Rao R, Sahin F (2002) Cluster-head identification in ad hoc sensor networks using particle swarm optimization. In: Proceedings of the IEEE international conference on personal wireless communications, New Delhi, pp 201–205
- Vinay Kumar SJ, Tiwari S (2011) Energy efficient clustering algorithms in wireless sensor networks: a survey. *Int J Comput Sci Issues* 8(5):259–268
- Xue F, Sanderson A, Graves R (2006) Multi-objective routing in wireless sensor networks with a differential evolution algorithm. In: 2006 IEEE international conference on networking, sensing and control, Ft. Lauderdale, FL, pp 880–885. <https://doi.org/10.1109/ICNSC.2006.1673263>

Fairness

Tian Lan

Electrical and Computer Engineering, George Washington University, Washington, DC, USA

Given a vector $\mathbf{x} \in \mathbb{R}_+^n$, where x_i is the resource allocated to user i , *how fair* is it? Consider two allocations among three users: $\mathbf{x} = [1, 2, 3]$ and $\mathbf{y} = [1, 10, 100]$. Among the large variety of choices for quantifying fairness, it is possible to have fairness values such as 0.33 or 0.86 for $\mathbf{x}$ and 0.01 or 0.41 for $\mathbf{y}$: $\mathbf{x}$ is viewed as 33 times more fair than $\mathbf{y}$, or just twice as fair as $\mathbf{y}$. How many such “viewpoints” are there? What would disqualify a quantitative metric of fairness? Can they all be constructed from a set of simple statements taken as true for the sake of subsequent inference?

Fairness of $\mathbf{x}$ can be quantified through a fairness measure, which is a function f that maps $\mathbf{x}$ into a real number. These measures are sometimes referred to as diversity indices in statistics. Various fairness measures have been proposed throughout the years. These range from simple ones, e.g., the ratio between the smallest and the largest entries of $\mathbf{x}$, to more sophisticated functions, e.g., Jain’s index and the entropy function. Some of these fairness measures map $\mathbf{x}$ to a normalized range between 0 and 1, where 0 denotes the minimum fairness, 1 denotes the maximum fairness (often corresponding to an $\mathbf{x}$ where all x_i

are the same), and a larger value indicates more fairness. For example, min-max ratio (Marson and Gerla 1982) is given by the maximum ratio of any two user’s resource allocation, while Jain’s index (Jain et al. 1984) computes a normalized square mean. How are these fairness measures related? Is one measure “better” than any other? What other measures of fairness may be useful?

An alternative method that has gained attention in the networking research community since (Kelly et al. 1998; Mo and Walrand 2000) is the optimization-theoretic approach of α -fairness and the associated utility maximization problem. Given a set of feasible allocations, a maximizer of the α -fair utility function satisfies the definition of α -fairness. Two well-known examples are as follows: a maximizer of the log utility function ($\alpha = 1$) is proportionally fair, and a maximizer of the α -fair utility function as $\alpha \rightarrow \infty$ is max-min fair. More recently, α -fair utility functions have been connected to divergence measures (Uchida and Kurose 2009). In Bonald and Massoulié (2001); Massoulié and Roberts (2002), the parameter α was viewed as a fairness measure in the sense that a fairer allocation is one that is the maximizer of an α -fair utility function with larger α – although the exact role of α in trading-off fairness and throughput can sometimes be surprising (Tang et al. 2006). While it is often believed that $\alpha \rightarrow \infty$ is more fair than $\alpha = 1$, which is in turn more fair than $\alpha = 0$, it remains unclear what it means to say, for example, that $\alpha = 3$ is more fair than $\alpha = 2$.

There are in general two main approaches to fairness evaluation: binary (is it fair according to a particular criterion, e.g., optimization objective, proportional fair, no-envy?) or continuous (quantifying how fair is it, and how fair is it relative to another allocation?). In the latter case, there are three sub-approaches:

- (A) A system-wide, global measure: (A1) $f(\mathbf{x})$ where f is our fairness function, or (A2) $f(U_1, U_2, \dots, U_n)$ where U_i is a utility function for each user that may depend on the entire $\mathbf{x}$.
- (B) Individual, global measures: the set of $\{f_i(\mathbf{x})\}$ (each user i cares about the entire allocation $\mathbf{x}$).
- (C) Individual, local measures: the set of $\{\tilde{f}_i(x_i)\}$ (user i only cares about her resource via some function $\tilde{f}_i$).

Renyi entropy. Renyi entropy is a family of functionals quantifying the uncertainty or randomness of generalized probability distributions, developed in 1960 (Renyi 1960). Renyi entropy is derived from a set of five axioms as follows:

- 1. Symmetry.
- 2. Continuity.
- 3. Normalization.
- 4. Additivity.
- 5. Mean-value property.

Renyi entropy, a generalization of Shannon entropy, has been further extended by others since the 1960s, e.g., smooth Renyi entropy for information sources (Renner 2005) and quantum Renyi entropy (Renner and Wolf 2004).

Lorenz Curve. Schur-concavity of fairness measures is a critical property for justifying fairness measures by establishing an ordering on the set of Lorenz curves (Aaberge 2001). Let $P_{\mathbf{x}}(y)$ be the cumulative distribution of a resource allocation $\mathbf{x}$. Its Lorenz curve $L_{\mathbf{x}}$, defined by

$$L_{\mathbf{x}}(u) = \frac{1}{\mu} \cdot \int_{\{P_{\mathbf{x}}(y) \leq u\}} y dP_{\mathbf{x}}(y), \quad (1)$$

is a graphical representation of the distribution of $\mathbf{x}$ and has used to characterize the social welfare distributions and relative income differences in economics. In 2001, an axiomatic characterization of Lorenz curve orderings is proposed based on a set of four axioms (Aaberge 2001):

- 1. Order. (The ordering is transitive and complete.)
- 2. Dominance. (The ordering is Shur-concavity.)
- 3. Continuity.
- 4. Independence.

Cooperative Economic Theories. In economics, a number of theories have been developed to study the collective decisions of groups. Many of these theories have also been uniquely associated with sets of axioms Moulin (1991), including two well-known axiomatic constructions: the Nash bargaining solution in 1950 (Nash 1950) and the Shapley value in 1953 (Shapley 1953).

The Nash bargaining solution was developed from a set of four axioms:

- 1. Invariance to affine transformation.
- 2. Pareto optimality.
- 3. Independence of irrelevant alternatives (IIA).
- 4. Symmetry.

Nash's axiom of IIA contributes most to his uniqueness result, and it is also often considered as a value statement. It has been shown by many others that replacing IIA with other value statements may result in solution classes different from the bargaining solution. Given a feasible region of individual utilities, the Nash bargaining solution is also equivalent to a maximization of the proportional fairness utility function.

Another well-known solution concept in economic study of groups is the Shapley value (Shapley 1953). In a coalition game, individual decides whether or not to form coalitions in order to increase the maximum utility of the group while ensuring that their share of the group utility is maximized. The Shapley value concerns an operator that maps the structure of the game, to a set of allocations of utility in the overall group.

Given a coalition game, four axioms uniquely define the Shapley value:

1. Pareto Optimality.
2. Symmetry.
3. Dummy.
4. Additivity.

Ultimatum Game. Cake-Cutting and Fair Division: In ultimatum game (Nowak et al. 2000), player A divides a resource into two parts: one part for herself and the other for player B. Player B can then choose to accept the division or reject it, in which case neither player receives any resource. Without a prior knowledge about player B's reaction, player A may divide anywhere between $[0.5; 0.5]$ and $[1; 0]$. Running this game as a social experiment in different cultures have led to debates about the exact implications of the results on people's perception on fairness: how fair does it take for player B's perception, and player A's guess of that, to accept player A's division? This has also been contrasted with the perception of fairness in the related dictator game (Ben-Ner et al. 2004; Bolton et al. 1998), where player B has no option but to accept the division by player A.

A classic generalization of ultimatum game that has received increasing attention in the past decade is the cake-cutting problem. As reviewed in books (Barbanel 2005; Brams and Taylor 1996), the cake is a measure space, and each player uses a countably additive, non-atomic measure to evaluate different parts of the cake. Among the work studying the cake-cutting problem, e.g., (Brams and Taylor 1995), the primary focus has been on two criteria: efficiency (Pareto optimality) and fairness (envy-freeness). Achievability results for four users or more are still challenging.

Fairness in cake-cutting and fair division is traditionally defined as envy-freeness. It is a binary summary based on each individual's local evaluation: the allocation of cake is fair if no user wants to trade her piece with another piece. In (Brams et al. 2006), this restrictive viewpoint on fairness is expanded to include proportional allocation of the left-over piece after each user gets $1/n$ th of

the cake (in her own evaluation). It is shown that Pareto optimality and proportional sense of fairness may not be compatible for three players or more.

Rawls' Theory of Justice and Distributive Fairness. In political philosophy, the work of Rawls has been both influential and provocative since the original publication in 1970 (Rawls 1971). Rawl's theory of justice concerns mostly with liberties, which are not exactly the same as resources. However, one may also read his theory as a qualitative framework for distributive fairness in allocating limited resources among users. The arguments posed by Rawls are based on two fundamental *principles* (axioms stated in English),

1. "Each person is to have an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others."
2. "Social and economic inequalities should be arranged so that they are both (a) to the greatest benefit of the least advantaged persons, and (b) attached to offices and positions open to all under conditions of equality of opportunity."

The first principle governs the distribution of liberties and has priority over the second principle. One may interpret it as a principle of distributive fairness in allocating limited resources among users. Next, the first part of Rawls second principle concerns with the distribution of opportunity, while the second part is the celebrated "difference principle": an approach different from strict egalitarianism (since it is on the absolute value of the least advantaged user rather than the relative value) and utilitarianism (when narrowly interpreted where the utility function does not capture fairness).

Normative Economics and Welfare Theory. Rawls theory also has intricate interactions with normative economics, where many results are analytic in nature [68]. In addition to stochastic dominance, Arrow's impossibility theorem, and the cooperation game theories of Nash and of

Shapley, there are several major branches (Atkinson 1970; Kolm 1969; Sen 1973). Another set is the ethical axioms of transfers. For example, the Pigou-Dalton principle states that inequality decreases via Robin Hood operation that does not reverse relative ranking. The principle of proportional transfers states that what the donor gives and the beneficiary receives should be proportional to their initial positions.

Bergson-Samuelson social welfare function (Dworkin 1981; Bergson 1938) $W(U_1(\mathbf{x}), \dots, U_n(\mathbf{x}))$ aims at enabling complete and consistent social welfare judgment on top of individual preference-based utility functions U_i . Kolm's theory of fair allocation (Kolm 1972) uses the criterion of equity as no-envy, and it is well-known that competitive equilibrium with equal budget is the only Pareto-efficient and envy-free allocation if preferences are sufficiently diverse and form a continuum (Varian 1976).

Sociology and Psychology: Inequality Indices.

Quantifying inequality/injustice/unfairness using individual, local measures has been pursued in sociology. For example, Jasso in 1980 (Jasso 1980) advocated justice evaluation index as log of the ratio between actual allocation and just allocation. Allocation can be done either in quantity or in quality (in which case ranking quantifies the quality allocation). Many properties were derived in theory and experimented with in data about income distribution in different countries (Jasso 2007). In particular, probability distribution of the index is induced by the probability distribution of the allocation. This index is derived based on two principles and three laws in the paper, including equal allocation maximizes justice, and aggregate justice is arithmetic mean of individual ones.

In Jasso (1999), two other injustice indexes, JI1 and JI2, were developed based on the above. One interesting feature is that JI1 differentiates between under-reward and over-reward as two types of injustice. Another useful feature is the decomposition of the total amount of perceived injustice into injustice due to scarcity and injustice due to inequality. They are further unified with Atkinson's measure of inequality (Atkinson

1970): 1 minus the ratio of geometric mean and arithmetic mean. At the heart of these indices is the approach of taking combinations of arithmetic and geometric means of an allocation to quantify the spread.

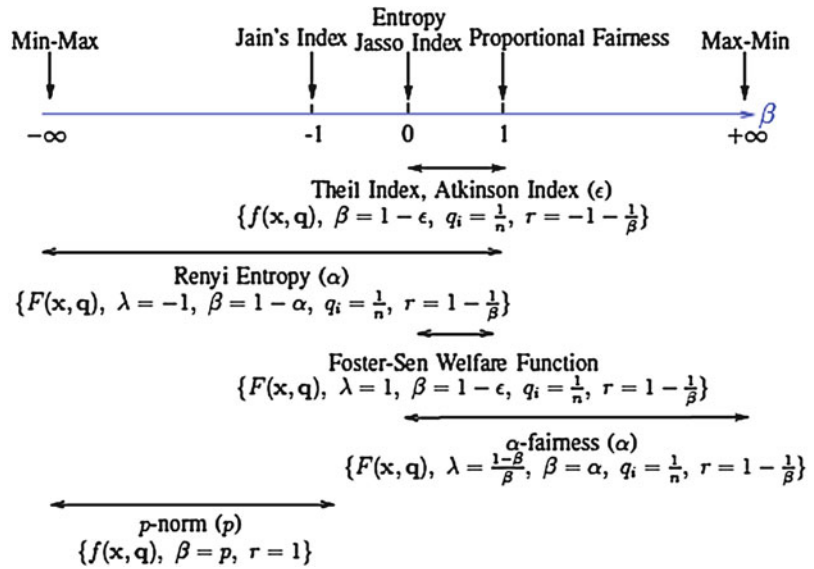
Lan-Chiang fairness theory. Many existing approaches for quantifying fairness are different. On the one hand, α -fair utility functions are continuous and strictly increasing in each entry of $\mathbf{x}$; thus, its maximization results in Pareto optimal resource allocations. On the other hand, scale-invariant fairness measures (ones that map $\mathbf{x}$ to the same value as a normalized $\mathbf{x}$) are unaffected by the magnitude of $\mathbf{x}$, and an allocation that does not use all the resources can be as fair as one that does. Can the two approaches be unified? To address the above questions, Lan and Chiang develop an axiomatic approach to fairness measures (Lan et al. 2010). It is shown that a set of five axioms, each of which simple and intuitive, thus accepted as true for the sake of subsequent inference, can lead to a useful family of fairness measures, i.e.:

1. Continuity.
2. Homogeneity.
3. Saturation.
4. Partition.
5. Starvation.

Starting with these five axioms, a unique family of fairness measures are generated from generator functions g : any increasing and continuous functions that lead to a well-defined "mean" function (i.e., from any Kolmogorov-Nagumo function Kolmogoroff 1930; Aczél 1948). Using power functions with exponent β as the generator function, a unique family of fairness measures are derived, i.e.:

$$f_{\beta, \lambda}(\mathbf{x}, \mathbf{q}) = \text{sign}(-1 - \beta) \cdot \left(\sum_i x_i \right)^{1/\lambda} \cdot \left[\sum_{i=1}^n q_i \left(\frac{x_i}{\sum_{j=1}^n x_j} \right)^{-\beta} \right]^{\frac{1}{\beta}}. \quad (2)$$

Fairness, Fig. 1 $F_{\beta,\lambda}$ unifies different fairness measures from diverse disciplines



This result unifies many existing fairness measures: Generalized Jain's index is a special case of $F_{\beta,\lambda}(\mathbf{x})$ for $1/\lambda = 0$ and $\beta < 1$; inverse of p -norm is another subclass of $F_{\beta,\lambda}(\mathbf{x})$ for $\beta \leq -1$; and α -utility is obtained for $1/\lambda = \beta/(1 - \beta)$ and $\beta > 0$. New fairness measures are revealed corresponding to other ranges of β and λ . The degree of homogeneity $1/\lambda$ determines how $F_{\beta,\lambda}(\mathbf{x})$ scales as throughput increases, while β provides trade-off between "resolution" and "strictness" of the fairness measure. The unification is illustrated in Fig. 1, including all known fairness measures that are global (i.e., mapping a given allocation vector to a single scalar) and decomposable (i.e., subsystems fairness values can be somehow collectively mapped into the overall systems fairness value).

References

- Aaberge R (2001) Axiomatic characterization of the Gini coefficient and Lorenz curve orderings. *J Econ Theory* 101:115–132
- Aczél J (1948) On Mean Values, *Bulltin of American Mathematics Society*, 54:392–400
- Atkinson AB (1970) On the measurement of inequality. *J Econ Theory* 2:244–263
- Barbanel J (2005) *The geometry of efficient fair division*. Cambridge University Press, Cambridge
- Ben-Ner A, Putterman L, Kong F, Magan D (2004) Reciprocity in a two-part dictator game. *J Econ Behav Organ* 53:333–352
- Bergson A (1938) A reformulation of certain aspects of welfare economics. *Q J Econ* 52:310–334
- Bolton GE, Katok E, Zwick R (1998) Dictator game giving: rules of fairness versus acts of kindness. *Int J Game Theory* 27:269–299
- Bonald T, Massoulié L (2001) Impact of fairness on internet performance. In: *Proceedings of ACM sigmetrics*
- Brams SJ, Taylor AD (1995) An envy-free cake division protocol. *Am Math Mon* 102:9–18
- Brams SJ, Taylor AD (1996) *Fair division: from cake-cutting to dispute resolution*. Cambridge University Press, Cambridge
- Brams SJ, Jones MA, Klamler C (2006) Better ways to cut a cake. *Not AMS* 53:1314–1321
- Dworkin R (1981) What is equality? Part 1: equality of resources. *Philos Public Aff* 10:185–246
- Jain R, Chiu D, Hawe W (1984) A quantitative measure of fairness and discrimination for resource allocation in shared computer system. *DEC Technical Report* 301
- Jasso G (1980) A new theory of distributive justice. *American Sociological Review* 45:3–32
- Jasso G (1999) How much injustice is there in the world? Two new justice indexes. *Am Sociol Rev* 64:133–158
- Jasso G (2007) Theoretical unification in justice and beyond. *Soc Justice Res* 20:336–371
- Kelly FP, Maulloo A, Tan D (1998) Rate control in communication networks: shadow prices, proportional fairness and stability. *J Oper Res Soc* 49:237–252
- Kolmogoroff A (1930) Sur la notion de la moyenne. *Atti della R. Accademia nazionale dei Lincei* 12
- Kolm SC (1969) The optimal production of social justice. In: Margolis J, Guitton H (eds) *Public economics*. Macmillan, London

- Kolm SC (1972) Justice and equity. MIT Press, Cambridge
- Lan T, Kao D, Chiang M, Sabharwal A (2010) An axiomatic approach to fairness in network resource allocation. In: Proceedings of IEEE INFOCOM
- Marson MA, Gerla M (1982) Fairness in local computing networks. In: Proceeding of IEEE ICC
- Massoulie L, Roberts J (2002) Bandwidth sharing: objectives and algorithms. *IEEE/ACM Trans Netw* 10(3):320–328
- Mo J, Walrand J (2000) Fair end-to-end window-based congestion control. *IEEE/ACM Trans Netw* 8(5):556–567
- Moulin H (1991) Axioms of cooperative decision making. Cambridge University Press, Cambridge
- Nash JF (1950) The bargaining problem. *Econometrica* 18:155–162
- Nowak MA, Page KM, Sigmund K (2000) Fairness versus reason in the ultimatum game. *Science* 289:1773–1775
- Rawls J (1971) A theory of justice. Harvard University Press, Cambridge
- Renner R (2005) Security of quantum key distribution. Ph.D. Dissertation at ETH, no. 16242, Sept 2005
- Renner R, Wolf S (2004) Smooth Renyi entropy and applications. In: Proceedings of IEEE ISIT
- Renyi A (1960) On measures of information and entropy. In: Proceedings of the 4th Berkeley symposium on mathematics, statistics and probability
- Sen AK (1973) On economic inequality. Clarendon Press, Oxford
- Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games, vol II. Annals of mathematical studies, vol 28. Princeton University Press, Princeton, pp 307–317
- Tang A, Wang J, Low S (2006) Counter-intuitive behaviors in networks under end-to-end control. *IEEE/ACM Trans Netw* 14(2):355–368
- Uchida M, Kurose J (2009) An information-theoretic characterization of weighted α -proportional fairness. In: Proceedings of IEEE INFOCOM
- Varian H (1976) Two problems in the theory of fairness. *J Public Econ* 5:249–260

Fault Tolerance and Reliability of Smart Grids

Sushmita Ruj¹ and Arindam Pal²

¹Indian Statistical Institute, Kolkata, India

²TCS Research and Innovation, Kolkata, India

Synonyms

[Robustness of smart grids](#)

Definitions

Electrical power grids are the backbone of the modern society. Since every aspect of our life depends on electricity, it is of utmost importance that the power grid infrastructure is highly reliable and always available. *Smart Grid* promises this by integrating the power grid and communication network. We can think of a smart grid as an interconnected network of power stations and communication centers. The interactions between these entities have important consequences on the resilience of smart grids. The failure of a small number of nodes can result in a cascading failure of a large number of nodes, which can bring the entire network down. Such grid failures have occurred in recent past in North America, Italy, and India. The effect of random and targeted attacks can lead to very different type of cascading failures. In this entry, we discuss different network models of smart grids and their impact on the reliability and availability. We believe that in order to design a robust smart grid, it is important to understand the interconnections amongst the different elements of the smart grid and the effect of one network on the other, under failure and attacks.

Introduction

A *smart grid* is an electrical power grid that uses information and communications technology (ICT) to automate the operations and to improve the efficiency, reliability, and sustainability of the production, transmission, and distribution of electricity. It is also sometimes referred to as an *intelligent power grid*.

The term smart grid was first used by Michael T. Burr (2003). These systems use bidirectional communication technology and computer processing that has been used for many years in other industries. They offer several benefits to utilities and consumers, resulting in big improvements in energy efficiency on the electricity grid and in homes and offices.

Smart grid integrates power grid and communication network. The power network consists of power plants, distribution stations, and electrical devices. A communication network consists of sensors to sense information, processing stations to process data and control stations. A key feature of the smart grid is the automation technology that lets the utility adjust and control each individual device or millions of devices from a central location.

Need for Designing Robust Smart Grids

Reliability of power grids is a very important consideration during its design and operation. In the USA, the annual cost of outages in 2002 is estimated to be \$79 billion, which is about a third of the total electricity revenue of \$249 billion.

In recent years, there have been some major blackouts in different parts of the world, affecting millions of people and causing losses to the tunes of billions of dollars to homes and businesses. We look at three such important blackouts. In each of these cases, the outage spreads a large area because of cascading failure of power grids.

1. **2003 blackout of USA and Canada:** The Northeast blackout of 2003 was a widespread power outage that occurred on Thursday, August 14, 2003, throughout parts of the Northeastern and Midwestern USA and the Canadian province of Ontario. At the time, it was the second most widespread blackout in history, after the 1999 Southern Brazil blackout. The blackout affected an estimated 10 million people in Ontario and 45 million people in 8 US states.
2. **2003 Italy blackout:** The 2003 Italy blackout was a serious power outage that affected all of Italy, except the islands of Sardinia and Elba for 12 h, and part of Switzerland near Geneva for 3 h on 28 September 2003. It was the largest blackout in the series of blackouts in 2003, affecting a total of 56 million people. It was also the most severe blackout in Italy in 70 years.
3. **2012 India blackout:** The July 2012 India blackout was the largest power outage in history, occurring as two separate events on July

30 and July 31 in 2012. The outage affected over 620 million people, about 9% of the world population, or half of India's population, and spread across 22 states in Northern, Eastern, and Northeast India. An estimated 32 gigawatts of generating capacity was taken offline in the outage.

Maintaining the reliability of power grids is becoming a major challenge due to the following reasons.

- Grid congestion due to uncertainty, diversity, and uneven distribution of energy supplies.
- Transmission of large amount of power over long distances increasing volatility and reducing reliability.
- Increasing energy consumption and higher level of peak demand.
- Aging infrastructure and less investment for maintenance.
- Maximizing utilization using modern tools for monitoring, analyzing, and control.
- Very high utilization of distributed resources increasing complexity and vulnerability of the grid.

Types of Faults in Smart Grids and Measures to Prevent Them

A smart grid is susceptible to various types of faults. These include overloading of grid, electrical and mechanical failures of generators and substations, sabotages, and terrorist attacks. Failures can occur because of incorrect routing algorithms.

In modern network topologies, message delivery failure may be avoided through smart routing technologies that can bypass faulty routers and links. Such fault tolerance is possible only when the faulty equipment does not constitute a single point of communication failure. Therefore, it is important to maintain redundant pathways through the network.

In order to build a reliable smart grid, it is important not only to identify the vulnerabilities but also study the structure of smart grid networks. Since a smart grid consists of millions of devices, it is important to understand how these

are interconnected. An attack/failure can affect one node, which in turn will affect other nodes and this process follows in a cascading way. In this entry, we therefore study the structure of the underlying smart grids. We model a smart grid as an interdependent complex network. Complex networks have been widely studied by computer scientists, mathematicians, physicists, biologists, and sociologists to model many real-world networks like social networks, biological networks, ad hoc networks, and citation networks (Pal and Ruj 2015). The complex network of power grid has been studied in great details (Pagani and Aiello 2013). However, the structure of smart grid is very different from that of a power grid, in that it consists of two networks which are interconnected. In order to build a robust smart grid, we therefore study the structure of individual networks, their interconnection and cascading failure between these two networks.

We model the structure of smart grids as interdependent networks in the section “[Structure of Interdependent Smart Grid Networks](#).” We discuss cascading failures in smart grids in the section “[Cascading Failures](#).” Attack models are presented in the section “[Attacks on Smart Grids](#)” along with measures of resilience. We look at techniques to increase resilience of networks in the section “[Improving the Resilience of Smart Grids](#).” We perform a comparative study of proposed models in the section “[A Comparative Study of Proposed Models](#)” and present experimental results. We conclude in the section “[Conclusion and Future Direction](#)” with some open problems and future directions.

The Smart Grid as an Interdependent Network

Before studying the structure of smart grid, we give a brief idea about complex networks.

Structure of Complex Networks

Networks occurring in real life have random structures and can be represented by some well-known graphical models. Degree distribution of

the graph G is the probability distribution of degrees of the vertices in G .

Probability that a node has degree k is calculated by dividing the number of nodes of degree k with the total number of nodes in the network. One of the earliest and simplest models is *Erdos-Renyi* (ER) random graph model. A graph $G(n, p)$ is called an Erdos-Renyi random graph if the graph consists of n vertices and every edge occurs with a probability p . We can also think of it as a graph $G(n, m)$, which is chosen uniformly at random from the set of all possible graphs on n vertices and m edges. The degree distribution of ER graph is Poissonian. For a large number of nodes, the degree distribution follows a normal distribution, which has a bell-shaped curve, when the degree k is plotted against the number of nodes of degree k .

Real-world networks like the Internet, World Wide Web, citation networks, and metabolic networks do not have a Poisson distribution (Pal and Ruj 2017). The degree distribution of such networks follows a power law distribution. In a power law degree distribution, the degree distribution is given by $P(k) \propto k^{-\alpha}$, where α is a constant ($1 \leq \alpha \leq 3$). Such networks are also known as scale-free (SF) networks. In such networks, there are many nodes with small degree and a very few high degree nodes. The high degree nodes are called hubs. The degree distribution decays slowly as the degree increases and gives rise to heavy tailed curve. Another class of networks has exponential degree distribution. Exponential degree distribution means that the probability that a node has degree k is given by $P(k) \propto e^{-\lambda k}$, where λ is a constant. The curve has a faster decay than power law distribution. Social networks exhibit *small-world* property. *Small-world networks* have small average path length between nodes.

From prior observations (Pagani and Aiello 2013), it is known that the power grid has either a power law degree distribution or exponential degree distribution. For example, the US power grid has exponential degree distribution, whereas the Italian power grid has power law degree distribution. The communication network follows a power law degree distribution according to the

results given in Huang et al. (2013), Dong et al. (2012) and Ruj and Pal (2014).

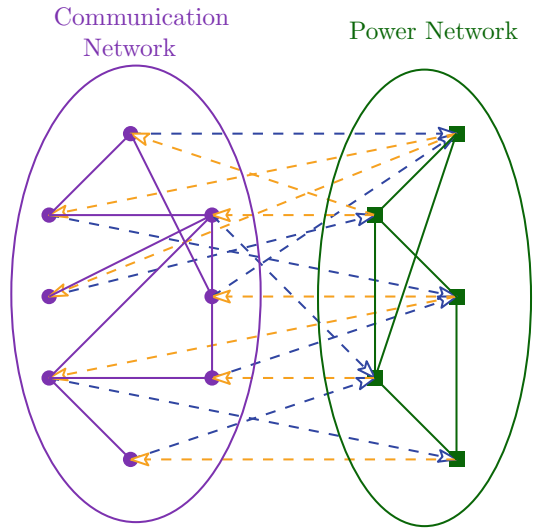
Structure of Interdependent Smart Grid Networks

An *interdependent network* is a network consisting of two subnetworks, where the nodes in each of the subnetworks depend on the nodes in the other subnetwork. The links across networks are called *dependency links* or *interlinks*, which are in addition to the links within the two subnetworks (also called *intralinks*). Since networks rarely appear in isolation, an interdependent network is a more realistic way to model network structures formed in real life. Interdependency between networks can explain events like the 2003 Italy blackout.

We represent a smart grid interdependent network as a graph $G_{SG} = (V_{SG}, E_{SG})$, consisting of two disjoint subnetworks represented by directed subgraphs $G_C = (V_C, E_C)$ and $G_P = (V_P, E_P)$. They denote the communication network and the power network, respectively. The number of nodes in the communication network and the power network are $|V_C| = n_C$ and $|V_P| = n_P$, respectively. The interlinks E_{PC} have one endpoint in V_C and the other endpoint in V_P . Thus, $E_{PC} = \{uv : u \in V_C, v \in V_P\} \cup \{uv : u \in V_P, v \in V_C\}$, and $E_{SG} = E_P \cup E_C \cup E_{PC}$.

There are different ways in which interlinks can be assigned between the nodes of the two networks. Power and communication nodes can be linked one-one, or can be linked many-one, or many-many. A power node must be controlled by at least one communication node, whereas a communication node must be powered by at least one power node. Depending on this, the in-degree of power and communication node both must be at least one. There can be added restrictions, where each communication node controls exactly one power node (out-degree of communication node is one), whereas, each power node supplies power to one or more communication nodes (out-degree of power node is at least one). The network structure and assignment of interlinks determine robustness.

In Fig. 1, we model the smart grid as an interdependent network of communication network



Fault Tolerance and Reliability of Smart Grids, Fig. 1
The smart grid as an interdependent complex network (Ruj and Pal 2014)

and power network. There are directed edges from communication nodes to power nodes and vice-versa. The links within each subnetwork are undirected edges. The structure of intralinks and interlinks together determine the resilience of the smart grid.

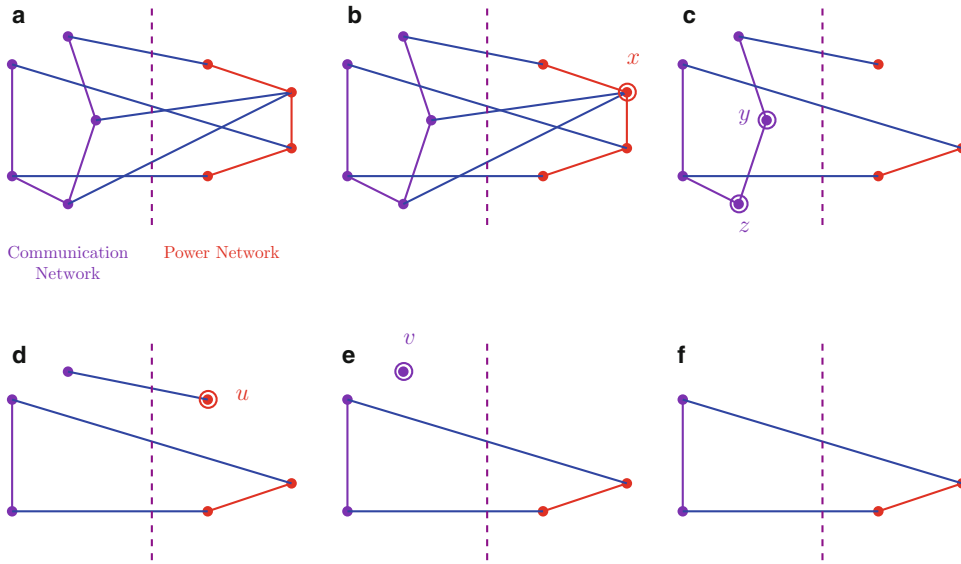
Fault Tolerance in Smart Grids

Cascading Failures

Electrical blackouts often occur in power grids because of cascading failures. In a *cascading failure* in an *interdependent network*, when a node fails in one network G_P , the nodes on the other network G_C which are dependent on this node also fail. Consequently, the nodes in G_P which are dependent on these nodes fail and the process continues recursively. Due to cascading failures, a large portion of a network can go down within a short amount time, even when a small fraction of nodes fail initially. This is a serious threat to the efficient operation of a network.

How Does It Occur?

We consider the example in Fig. 2. Initially, the node x in the power network G_P fails. As a result,



Fault Tolerance and Reliability of Smart Grids, Fig. 2 Illustration of cascading failure in interdependent smart grid networks. The faulty nodes are shown as double circles

the nodes y and z in the communication network G_C fail, because they lose the interlink from node x . Next, node u in G_P fails, as it becomes an isolated node. Then, node v in G_C fails as it becomes an isolated node. The cascading failure stops at this point, and the final network is shown in Fig. 2f.

Attacks on Smart Grids

A smart grid can be attacked in many different ways. Here, we are assuming an adversarial model, meaning that there is an adversary who is trying to cause node and/or edge failures to disrupt the functioning of the network. There can be many cybersecurity attacks on smart grids, like denial-of-services (DoS) attacks, network intrusion, compromising communication equipments, and interception. The details can be found in Mo et al. (2012). Here we consider node compromise under random, targeted, and adaptive attacks.

Random Attacks

In a random attack, the adversary selects a vertex/edge uniformly at random from the collection of all possible vertices/edges and makes it faulty. When a vertex becomes faulty, the vertex along

with all the edges incident on it is removed from the network.

Targeted Attacks

In a targeted attack, the adversary selects a vertex with a probability proportional to its degree (or some other parameter such as betweenness centrality or clustering coefficient) and makes it faulty. Our intuition suggests that if high-degree nodes are removed, the network disintegrates faster than if nodes are removed at random. Betweenness of a node is the number of shortest paths that pass through the nodes. Compromising a high-degree node disrupts a large number of paths between nodes, thus disconnecting them. High clustering coefficient means that the neighbors of a node are highly likely to be neighbors themselves.

Adaptive Attacks

In adaptive attacks, the attacker deletes nodes iteratively in rounds, instead of all nodes at once. In each round, the attacker chooses a set of nodes to be compromised based on the new centrality measure and compromises this set. It has been observed (through experiments) that an adversary has an advantage while compromising nodes adaptively, compared to nonadaptive deletion.

Measuring Robustness of Smart Grid Networks

Having modeled a smart grid as a complex interdependent network, we are now in a position to define the different measures of robustness. One way of doing this is to find the fraction of nodes in the *giant component* that survived the attack. The giant component is a connected component having a constant fraction of nodes of the original network. In other words, if there are n nodes in the network, a giant component has $O(n)$ nodes. Robustness can thus be calculated by finding the size of the giant component, if it exists. If a giant component does not exist after an attack, then we can say that the network is not resilient to attacks.

We can also measure robustness by the fraction of nodes that can be attacked, in order to maintain a giant component. It has been observed that there is a threshold at which a network crumbles down. This means that if we compromise fewer nodes, then there will be a giant component. However, above this value, the network disintegrates. The fraction of nodes surviving when transition takes place is called the *critical probability*.

Though we want nodes to connect to each other using small hops, it might so happen that after an attack, two nodes are still connected by a very long path. Connectivity is preserved, but the distance between nodes becomes substantially large. Resilience can be measured by studying the size of the giant component of small diameter, if there exists any.

Another measure is the average path length in a network. The *average path length* is the shortest path length between two nodes averaged over all possible node pairs. The shorter the average path length, the better connected the network is, and hence more fault tolerant.

Improving the Resilience of Smart Grids

The most important question is to decide the structure of the network, which will make the smart grid robust. One way to achieve this is to find a proper assignment of interlinks. One might think that increasing the number of interlinks

might increase the robustness. However, if nodes of high degree are compromised, the smart grid might collapse. The cost of maintaining large number of interlinks is also very high. Thus, on one hand we have the problem of increasing robustness, and on the other hand is the problem of cost. So, we need to have a trade-off between the two.

Apart from increasing the number of interlinks, we can change the structure of the networks. The current literature (Ruj and Pal 2014; Dong et al. 2012) suggests the use of scale-free (SF) networks for both power and communication. One reason is that, some power grids are known to be scale-free. However, there are no reasons to believe that SF-SF combination will result in better resilience to attacks. As shown in Ruj and Pal (2014), SF networks are more prone to targeted attacks than ER networks. The question is then, what should be the structure of the power network and communication network, which will guarantee high resilience in case of both random and targeted attacks.

There has been some work on mitigating attacks in power networks (Schneider et al. 2011; Wang et al. 2012). These papers study the structure of networks that improves resilience. For example, Schneider et al. (2011) increases the resilience of networks by swapping edges. Suppose there are two edges ab and cd . Edges are swapped, resulting in ac and bd . If this swapping increases the resilience, then this configuration is considered. Else, the old configuration is preserved. This results in the network structure which looks like an onion, with high-degree nodes at the center and low-degree nodes at the periphery.

Apart from making structural changes to networks, we can also look at easy ways to identify and stop cascading failures. A recent paper Kaplunovich and Turitsyn (2013) presents an algorithm which identifies dangerous pairs of failures in power grids and prunes them in an efficient manner. The algorithm performs very well for medium size power networks consisting of about 3000 nodes and prunes 99% of failure nodes in just 10 min. Ruj and Pal (2016) initiated the study of preferential attachment model with degree bound. They apply this idea to the key

predistribution problem in wireless sensor networks (WSN) and smart grids.

Huang et al. (2015) studied the effect of cascading failures in interdependent complex networks (smart power grids, automated traffic control systems, wireless sensor, and actuator networks) using percolation theory. The physical resource and computational resource networks are connected and mutually dependent. The failure in the physical resource network might cause failures in the computational resource network, and vice versa. A small failure in either of them could trigger cascade of failures within the entire system. They did a thorough mathematical analysis of failure propagation in the system. Their mathematical analysis proves that there exists a threshold for the proportion of faulty nodes, above which the system collapses. Using extensive simulations, they determine the critical values for different system parameters.

A Comparative Study of Proposed Models

We now study the existing models of smart grids that have been proposed so far. We present them in Table 1. The table describes the type of networks that have been considered so far, the interconnection patterns, the attack models and countermeasures, and the important contributions. Most of the papers (Huang et al. 2013, Dong et al. 2012; Ruj and Pal 2014; Parandehgheibi and Modiano 2013) consider smart grids exclusively; (Yagan et al. 2012) discusses cyber-physical systems with emphasis on smart grids, (Schneider et al. 2011; Wang et al. 2012) discuss power grids. Most of the papers consider the same type of networks, ER-ER or SF-SF. Buldyrev et al. (2010) initiated the study of interdependent networks. They considered random attacks on networks of same size and one-one assignment of interlinks. They pointed out that SF-SF interdependent networks

are less robust than single SF networks. Yagan et al. (2012) studied ER networks, where each node has the same inter-degree. They show that such a regular allocation of links is more robust to failures than random assignment of links. Huang et al. (2013, Dong et al. 2012) considered smart grids with SF-SF power-communication networks and studied the percolation threshold. They conducted extensive experiments by random node compromise to show that the estimated percolation threshold depends on the power law exponent. They also gave estimates of cost vs. robustness. Adding more interlinks increases robustness but is costly. Ruj and Pal (2014) studied targeted attacks on smart grids and showed that the SF-SF networks are more prone to targeted attacks than ER-ER networks. Compromising a small fraction of high-degree nodes disintegrates the network, i.e., it does not have a giant component.

The problem of interdependent smart grids has been modeled in (Nguyen et al. 2013; Sen et al. 2014) as layered graphs with power and communication graphs at the two layers. In Nguyen et al. (2013), the authors have considered the problem of finding the set of nodes that has to be removed to minimize the number of nodes in the largest connected component. This problem is called *Interdependent Power Node Disruptor* (IPND), and it has been shown to be NP-Hard. They show that this problem cannot be approximated better than a factor of $(2 - \epsilon)$ to the optimum. They have given a greedy algorithm to solve IPND. Sen et al. (2014) have shown that there exist polynomial time algorithms for the above problem in special cases of interconnection. One such case is when a power node depends on a single communication node and vice versa.

Parandehgheibi and Modiano (2013) studied a simple version of smart grids in which all power nodes are connected to a main hub and all communication nodes are connected to a main control center, thus forming a star-like structure. They proposed an algorithm to identify the minimum number of nodes that will disconnect the whole network.

Fault Tolerance and Reliability of Smart Grids, Table 1 Comparison of existing works

References	Network type	Network size	Type of interlinks	Attack type	Countermeasures	Important contribution
Buldyrev et al. (2010)	ER-ER/SF-SF	$n_C = np$	1–1	Random	None	Initiated the study of interdependent networks
Yagan et al. (2012)	ER-ER	$n_C = np$	Regular allocation [†]	Random	None	Regular allocation performs better than random allocation
Huang et al. (2013)	SF-SF	$n_C > np$	Random allocation	Random	None	Finding critical probability
Dong et al. (2012)	SF-SF	$n_C > np$	Random allocation	Random	None	Studying cost vs. resilience
Ruj and Pal (2014)	SF-SF	$n_C > np$	Random allocation	Targeted	None	SF networks are more susceptible to targeted attacks than ER networks
Nguyen et al. (2013)	Any	Any	Random	Targeted	None	Finding a set of vulnerable nodes minimizing the largest connected component is NP-Hard
Sen et al. (2014)	Any	Any	Random	Targeted	None	Algorithms for finding vulnerable set maximizing the size if the disconnected vertices
Parandehgheibi and Modiano (2013)	Star graphs	Any	Random allocation	Targeted	None	Smallest set of nodes to be revoked to disconnect the whole network
Schneidera et al. (2011)	SF networks	Any	None	Targeted	Onion	Small changes in structure can vastly affect robustness
Wang et al. (2012)	SF networks	Any	None	Targeted node/link	Onion structure	Proposed new measures of robustness
Dong et al. (2013)	Network-of networks	Any	Tree of networks	Targeted	None	Robustness depends on connection pattern of networks

Experiments on Interdependent Networks and Implications

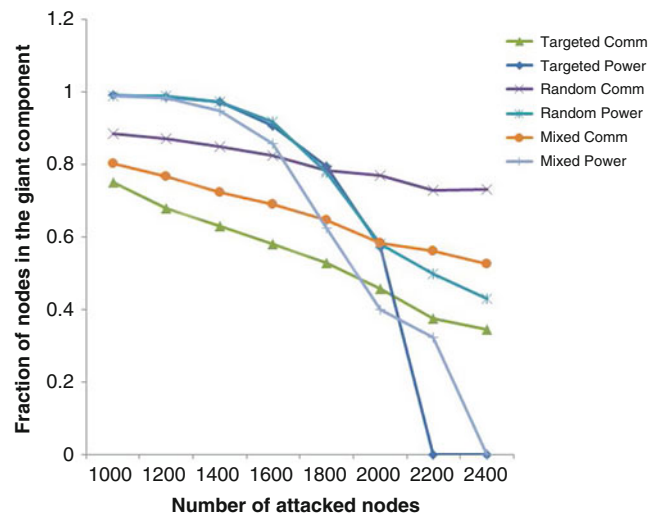
In this section, we look at experimental results on interdependent smart grids. We consider different types of interdependent networks like ER-ER and SF-SF, of different sizes and different interconnection patterns. We study random and targeted attacks on these networks. We show that SF-SF graphs are more vulnerable to targeted attacks than ER-ER graphs. The experiments have been performed using the network library *igraph* (Csardi and Nepusz 2006) on C. *igraph* has inbuilt packages for performing different functions like finding giant components, average path lengths, and betweenness.

In Fig. 3, the power network consists of 1000 nodes and the communication network consists of 10,000 nodes, following the conventions of Ruj and Pal (2014) and Dong et al. (2012). The communication/power network is generated as a scale-free network using a power-law degree distribution. Interlinks have been assigned randomly from communication to power network. We have plotted the size of the giant component (as a fraction of the size of the communication/power network) against the number of nodes attacked in the communication network. In targeted attacks, high-degree nodes are compromised with higher probability. We have considered three types of attacks on communication network: random, targeted, and mixed (50% nodes compromised ran-

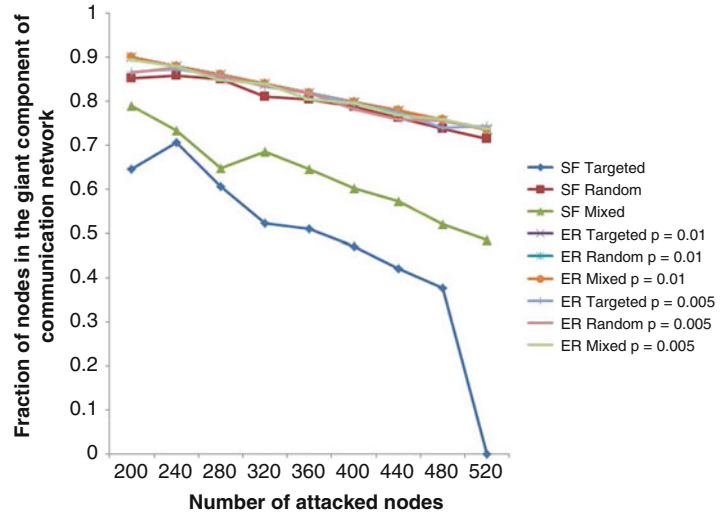
domly and 50% compromised based on their degree). It is observed that for a given value of the number of attacked nodes, the fraction of nodes in the giant component of the communication network is highest for random attacks and lowest for targeted attacks. The corresponding fraction for mixed attacks lies somewhere in the middle. We also see that for the same fraction of nodes compromised, the giant component of the power network disintegrates faster for targeted attacks, compared to random attacks. We see that on compromising 2200 nodes, there is no giant component when targeted attack occurs, whereas giant component exists, under random attacks. This is expected, as attacking higher degree nodes result in a faster disintegration of the network, resulting in smaller components. A similar result is expected when we consider betweenness instead of degree, i.e., compromise nodes of high betweenness compared with high probability.

In Figs. 4 and 5, the communication network consists of 2000 nodes, whereas the power network consists of 1000 nodes. In Fig. 4, we have plotted the size of the giant component (as a fraction of the size of the communication network) against the number of attacked nodes, where as, in Fig. 5, we have plotted the size of the giant component (as a fraction of the size of the power network) against the number of attacked nodes. The communication network and power networks

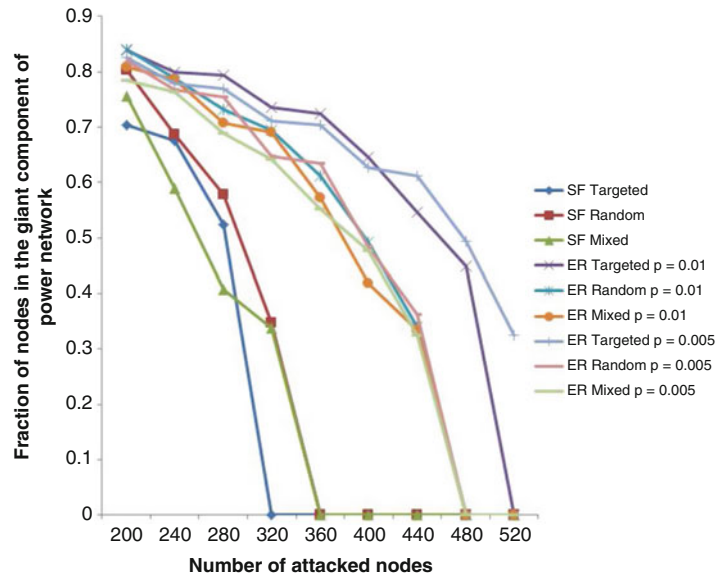
Fault Tolerance and Reliability of Smart Grids, Fig. 3 Variation of giant component size with number of attacked nodes for targeted, random, and mixed attacks



Fault Tolerance and Reliability of Smart Grids, Fig. 4 Variation of giant component size with number of attacked nodes in the communication network for scale-free and Erdos-Renyi models



Fault Tolerance and Reliability of Smart Grids, Fig. 5 Variation of giant component size with number of attacked nodes in the power network for scale-free and Erdos-Renyi models



are generated using (1) a scale-free (SF) network using a power-law degree distribution, (2) the Erdos-Renyi (ER) $G(n, p)$ model with $p = 0.01$, and (3) the Erdos-Renyi (ER) $G(n, p)$ model with $p = 0.005$. Only nodes in the communication network are attacked. In Fig. 5, we have plotted the size of the giant component (as a fraction of the size of the power network) against the number of attacked nodes. From Figs. 4 and 5 we see that, for all the three types of attacks, the sizes of the giant components for ER graphs

are comparable. The power and communication networks will be more fault tolerant to targeted attacks for Erdos-Renyi networks, compared to scale-free networks.

Conclusion and Future Direction

In this entry, we discussed how power grids can become faulty and the different ways in which they can be attacked. We modeled the smart grid

as an interdependent network of a communication network and a power network. In future, the smart grid will be connected to the Internet of Things (IoT), so there might be more than two networks to consider. We studied some real-world scenarios where cascading of faults led to a catastrophic failure of a large portion of the smart grid. We discussed ways to measure the robustness of a smart grid. We also considered various strategies to design the networks so that the effect of attacks can be effectively neutralized. The structure of the communication network and the power network, and the interconnection pattern between the two can have a profound effect on the reliability of a smart grid. We need to study further to have a better understanding of the network structure on the fault-tolerance of a network.

Cross-References

- [Cyber-Physical Security in Smart Grid Communications](#)
- [Situation Awareness in Smart Grids](#)

References

- Buldirev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. *Nature* 464:1025–1028
- Csardi G, Nepusz T (2006) The igraph software package for complex network research. *Int J Complex Syst* 1695:2804. <http://igraph.sourceforge.net/>
- Dong G, Gao J, Tian L, Du R, He Y (2012) Percolation of partially interdependent networks under targeted attack. *Physical Review E* 85(1):6112
- Dong G, Gao J, Du R, Tian L, Stanley HE, Havlin S (2013) Robustness of network of networks under targeted attack. *Phys Rev E* 87:052,804. <https://doi.org/10.1103/PhysRevE.87.052804>
- Huang Z, Wang C, Ruj S, Stojmenovic M, Nayak A (2013) Modeling cascading failures in smart power grid using interdependent complex networks and percolation theory. In: The 8th IEEE conference on industrial electronics and applications, ICIEA13, pp 1023–1028
- Huang Z, Wang C, Stojmenovic M, Nayak A (2015) Characterization of cascading failures in interdependent cyber-physical systems. *IEEE Trans Comput* 64(8):2158–2168
- Kaplunovich PA, Turitsyn KS (2013) Fast selection of n^2 contingencies for online security assessment. In: Proceedings of 2013 IEEE power and energy society general meeting (PES)
- Mo Y, Kim THJ, Brancik K, Dickinson D, Lee H, Perrig A, Sinopoli B (2012) Cyber-physical security of a smart grid infrastructure. *Proc IEEE* 100(1): 195–209
- Nguyen DT, Shen Y, Thai MT (2013) Detecting critical nodes in interdependent power networks for vulnerability assessment. *IEEE Trans Smart Grid* 4(1): 151–159
- Pagani GA, Aiello M (2013) The power grid as a complex network: a survey. *Physica A* 392:26882700
- Pal A, Ruj S (2015) CITEEX: a new citation index to measure the relative importance of authors and papers in scientific publications. In: 2015 IEEE international conference on communications (ICC), IEEE, London, pp 1256–1261
- Pal A, Ruj S (2017) A graph analytics framework for ranking authors, papers and venues. *arXiv preprint arXiv:170800329*
- Parandehgheibi M, Modiano E (2013) Robustness of interdependent networks: The case of communication networks and the power grid. In: IEEE Globecom, Atlanta, GE
- Ruj S, Pal A (2014) Analyzing cascading failures in smart grids under random and targeted attacks. In: 28th IEEE international conference on advanced information networking and applications, AINA, Victoria, BC
- Ruj S, Pal A (2016) Preferential attachment model with degree bound and its application to key predistribution in WSN. In: 2016 IEEE 30th international conference on advanced information networking and applications (AINA), IEEE, Crans-Montana, Switzerland, pp 677–683
- Schneidera CM, Moreira AA, Andrade JS, Havlin S, Herrmann HJ (2011) Mitigation of malicious attacks on networks. *PNAS* 108(10):3838–3841
- Sen A, Mazumdar A, Banerjee J, Das A, Compton R (2014) Identification of k most vulnerable nodes in multi-layered network using a new model of interdependency. *CoRR*. <http://arxiv.org/abs/1401.1783>
- Wang W, Xu Y, Khanna M (2012) Enhancing network robustness against malicious attacks. *Phys Rev E* 85:066,130
- Yagan O, Qian D, Zhang J, Cochran D (2012) Optimal allocation of interconnecting links in cyber-physical systems: Interdependence, cascading failures and robustness. *CoRR abs/1201.2698*

Fingerprint-Based Positioning

- [Fingerprinting Localization](#)

Fingerprinting Localization

Zijun Gong¹, Fan Jiang¹, Kun Hao^{2,3}, and Yan Zhang^{2,3}

¹Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

²Department of Engineering and Applied Science, Memorial University of Newfoundland, St John's, NL, Canada

³Tianjin Chengjian University, Tianjin, China

Synonyms

Fingerprint-based positioning; RSSI-based positioning; Wi-Fi localization

Definition

Fingerprinting localization refers to a series of positioning techniques for localizing and tracking radio frequency (RF) devices by employing received signal strength (RSS) or received signal strength indicator (RSSI). They generally contain two phases: online phase and offline phase. In the first phase, RSS values of different beacons are measured on different reference locations in the area of interest. These values will be stored in a database. In the second phase, the target RF device can measure an RSS vector and compare this vector to those in the database. The positioning result will be the coordinate of the closest RSS vector or the weighted coordinate of several closest RSS vectors.

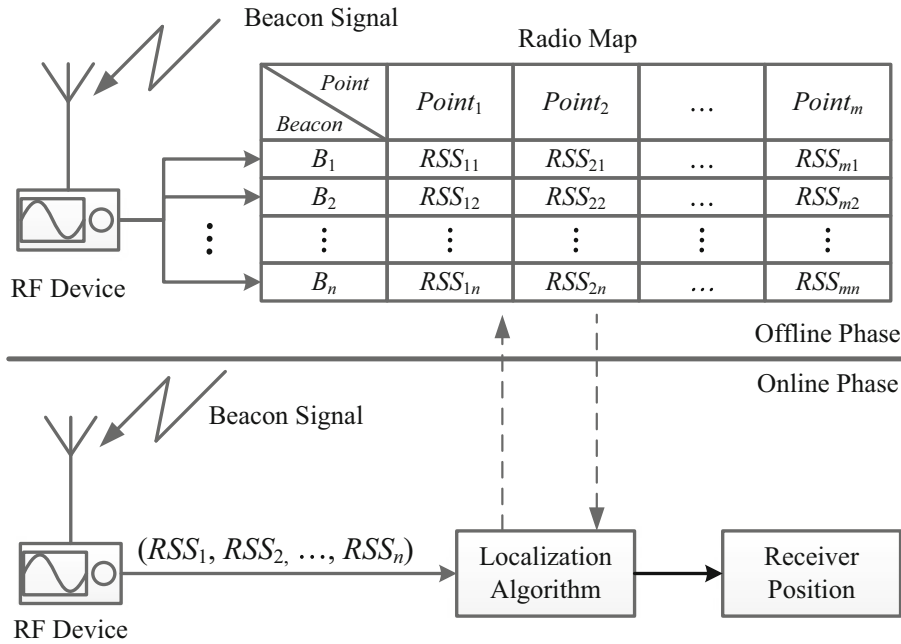
Historical Background

GPS has been a very successful positioning system for several decades. However, this system is not available in some scenarios, e.g., indoor environments. As alternatives to GPS, many different localization techniques have been proposed,

based on time of arrival (ToA), time difference of arrival (TDoA), angle of arrival (AoA), and RSS (RSSI). The first two metrics demand time synchronization among beacons and target devices, while AoA-based method requires cumbersome antenna arrays. Compared with these competitors, RSS-based system has significant advantages: (1) WLANs are deployed almost everywhere; (2) RSS measurements are readily available in most RF chips; and (3) the localization algorithms have very low complexity.

In 2000, the first fingerprint-based localization system named as RADAR was developed by Microsoft Research Asia (Bahl and Padmanabhan 2000), as shown in Fig. 1. In RADAR, some reference points are chosen across the area of interest. On every reference point, an RF device is employed to measure the RSS values from different beacons. These RSS vector is recorded in a database with the corresponding coordinate of the reference point. In other words, every coordinate corresponds to a RSS vector. This vector is unique for every coordinate and thus is also referred to as fingerprint of that spot. When a target device is located in the area of interest, it will first measure the RSS vector in its communication range and then compare this vector with those in the radio map. After finding the RSS vector with the least Euclidean distance, the corresponding coordinate is chosen as the localization result. This algorithm is often referred to as the nearest neighbor (NN) algorithm. The intuitive idea behind this method is that if two positions are physically closer, their RSS vectors should have smaller Euclidean distance. This is not always the case, but it is statistically correct.

In 2001, the authors proposed another Wi-Fi location service that uses Bayesian networks to infer the location of a device, called Nibble (Castro et al. 2001). In the radio map, the joint distribution of RSS values from different beacons at a specific point is recorded as an entry. Then, during online phase, the likelihood probabilities of the received RSS vector conditioned on the assumption that the target is located on reference points are calculated and compared. Then, the reference points with largest probabilities are



Fingerprinting Localization, Fig. 1 A typical fingerprinting-based localization system contains online and offline phases

assumed to be neighbors of the target, and the target's location is the weighted average of the neighbors. The intuitive idea is that if the target is close to a specific reference point, their RSS vectors should be similar, and the corresponding probability should be large.

Since RADAR and Nibble, many other fingerprinting-based localization systems and techniques have been developed, and most of them are variants of either RADAR or Nibble. Compared with their ancestors, the variants are adjusted for specific scenarios to achieve better accuracy and efficiency.

Foundations

There are generally two metrics, the Euclidean distance and maximum likelihood ratio estimation. For the first metric, the Euclidean distance between the real-time RSS vector and those in the radio map needs to be calculated. Apart from Euclidean distance, probability can also be employed for localization (Castro et al. 2001). In training phase, the empirical distribution of RSS

vectors is obtained by measurements. During offline phase, the likelihood probabilities of different positions will be calculated. The larger the probability is, the closer the physical distance between the target and the reference point will be.

For both metrics, NN and k -NN algorithms can be employed for final localization. The k -NN algorithm provides a weighted result of the k nearest neighbors in terms of Euclidean distance or probability. The weights can be normalized reciprocal of Euclidean distance or the probabilities.

In spite of various strong aspects of the fingerprint-based localization techniques, it still has several major drawbacks in practical applications, and possible solutions have been proposed.

The first problem is the significant labor work coming with the construction of the radio map. As a matter of fact, the density of reference points is important for localization accuracy. However, to obtain and maintain a high-resolution radio map is not easy. To be specific, the typical density of radio maps in indoor environments is

1 point per square meter and 100 samples per testing point, so as to average out the fluctuations in RSS measurements. Besides, the radio maps require constant update due to indoor environment change. There are several possible ways to solve this problem. The first strategy is to only sample RSS vectors on a portion of reference points and estimate the RSS vectors on other points based on radio propagation model. This strategy was employed by RADAR, and the radio propagation model was discussed in many literatures (Seidel and Rappaport 1992; Han et al. 2015). The second strategy is to employ robots to do the labor work. These two strategies both show promising results in certain environments, but their applications to complex indoor environments are quite limited.

The second problem is how to effectively find the nearest neighbor or neighbors in large radio maps. For real-time localization, the delay introduced by searching radio maps can be quite significant. One possible solution is to divide the entries in the radio map into clusters (Youssef et al. 2003). Intuitively, when reference points are close, their RSS vectors should have similar characteristics.

The last problem is often referred to as device diversity or device heterogeneity. To be specific, the RF chip of the target device generally is generally different from that used for radio map construction during offline phase. For the same received signal, different RF chips may report different RSS (RSSI) measurements. Therefore, we cannot just use Euclidean distance to compare the similarity of two RSS vectors. As alternatives, cosine similarity (He and Chan 2014) and Tanimoto similarity (Jiang et al. 2012) are better choices, and they have been implemented in practical applications and results are reported.

Key Applications

Fingerprint localization techniques can be employed for positioning in ad hoc networks in various environments.

Cross-References

► [RSSI-Based Positioning](#)

References

- Bahl P, Padmanabhan VN (2000) RADAR: an in-building RF-based user location and tracking system. In: Proceedings of the 9th annual joint conference of the IEEE computer and communications societies, Israel
- Castro P et al (2001) A probabilistic room location service for wireless networked environments. In: Proceedings 3rd international conference on ubiquitous computing, Georgia
- Han S et al (2015) An indoor radio propagation model considering angles for WLAN infrastructures. *Wirel Commun Mob Comput* 15(16):2038–2048
- He S, Chan G (2014) Sectjunction: Wi-Fi indoor localization based on junction of signal sectors. In: Proceedings of the 2014 IEEE international conference on communications, Sydney
- Jiang Y et al (2012) ARIEL: automatic Wi-fi based room fingerprinting for indoor localization. In: Proceedings of the 2012 ACM conference on ubiquitous computing, Pittsburgh
- Seidel SY, Rappaport TS (1992) 914 MHz path loss prediction models for indoor wireless communications in multifloored buildings. *IEEE Trans Antennas Propag* 40(2):207–217
- Youssef MA et al (2003) WLAN location determination via clustering and probability distributions. In: Proceedings of the 1st IEEE international conference on pervasive computing and communications, Texas

First-Order Reaction

► [Reaction-Diffusion Channels](#)

5G

► [Big Data in 5G](#)

5G NR

► [5G Wireless](#)

5G Wireless

Jin Yang

Verizon Communications, Walnut Creek, CA,
USA

Synonyms

3GPP release 15; 5G NR; 5G wireless; New radio (NR)

Definitions

5G wireless is the fifth generation cellular mobile technology based on specifications defined in Release 15 of 3GPP standards [1][2][3][4][5].

Historical Background

5G wireless is fundamentally transforming a radio network from pure wireless connectivity to the network for services. Mobile wireless access technologies have gone through several generations of evolutions to increase radio access capacity. The spectral efficiency has approaching Shannon capacity. However, there is enormous opportunity on support of various services. 5G wireless (3GPP TS 38.201-36.215 2017; 3G3GPP TR 38.801 2017; 3GPP TR 38.802 2017; 3GPP TR 38.803 2017; 3GPP TR 38.804 2017) is expected to support a variety of services more efficiently.

5G applications can range from Gigabit Society with personalized ultra-broadband TV and smartphone to autonomous car, connected home and cities, as well as remote sensors for agriculture and smart utility grid. It will dramatically improve our daily life. One example is healthcare. It could shift from curing to prevention. An ambulance could stop by your side at the moment you have a stroke or life-threatening event. Verizon *State of the IoT Market 2017* report has confirmed the double digital growth and expected game change from 5G services (Verizon 2017; Yang et al. 2016).

5G wireless will enable new services and applications, in particular, enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC). Figure 1 illustrated a service-based radio access architecture. The network is centrally controlled by an end-to-end self-organizing network (SON) with software-defined networking (SDN) in a virtualized radio access network (vRAN) and network function virtualization (NFV) environment.

This network transformation can support network slicing and differentiate services at access and transport nodes to ensure end-to-end performance, enabling fast service delivery. It gives us the ability to steer traffic in fine granularity to enforce policy-based routing and meet end-to-end Quality of Service (QoS) requirements of various applications.

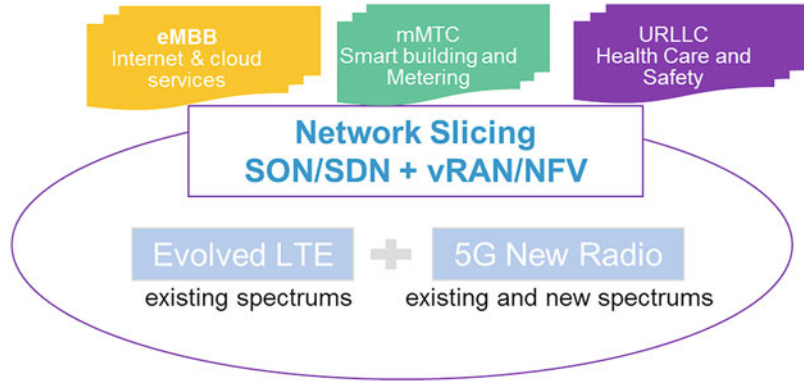
5G wireless consists of an evolution of today's 4G LTE (Long-Term Evolution) network as well as a new standardized radio access technology known as 5G New Radio (NR). The first standardization phase of NR is expected to be completed in 2018, with a sequence of releases following after. 5G wireless brings us opportunities for more innovative services and revenue streams and also more challenges to support those services efficiently and reliably.

Spectrum Management

5G wireless demands large bandwidth to achieve the gigabit per second (Gbps) user data rate. 5G NR can operate in the frequency ranging from hundreds of MHz to 100 GHz with paired and unpaired spectrum at more than 100 MHz bandwidth to achieve 10 Gbps targeted peak data rate. Field trials for 5G NR critical technologies are conducted at 4 GHz, as well as 27 GHz and 37 GHz. Thus, spectrum and resource management are extremely important.

The low-band in sub-6 GHz can be utilized for wide area services, while high-band for ultra-dense small cells or fixed wireless applications. 5G wireless needs to support licensed spectrum

5G Wireless, Fig. 1 Radio access network to support various services



F

and leverage unlicensed and shared spectrum to increase throughputs and complement user performance.

Millimeter wave (mmWave) with extensive hundreds of MHz is critical for average user throughput larger than 1 Gbps in a loaded dense network with cell radius less than 100 m. mmWave and relies on massive MIMO (multiple inputs multiple outputs) to mitigate sensitivity to rain and fogs. Massive antenna system can form and track user-specific narrow beam to ensure service quality and reliability. This can mitigate a relatively large propagation loss and recover from blockages.

Dual connectivity will gracefully extend initial 5G to existing LTE wide area coverage. 5G could use licensed assisted access scheme similar to LTE to leverage more than 500 MHz unlicensed spectrum under 6 GHz and even more in mmWave.

Idle and connected mode mobility management is essential to ensure consistent user experience across the network. Those are challenging in a heterogeneous network with large range of inter-site distances and frequency bands. NR supports mobility without reconfiguration and make-before-break handover with multiple connections. Thus NR reduces the handover interruption time to zero from current LTE's minimal 25 ms.

5G will bring in more flexibility on frequency and time allocation and allow flexible uplink and downlink multiplex to improve spectrum utilization. The 5G NR in Rel.15 will support more flexible Time Division Duplex (TDD) resource

allocations; hence each user can have individually specified downlink and uplink time slots. In later releases, we are expecting even more flexible full dynamic Frequency Division Duplex (FDD) and TDD allocation, as well as full duplex radios to improve Integrated Access Backhaul.

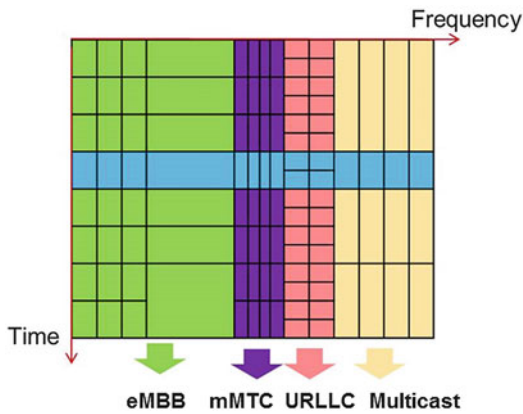
Fundamental of 5G Wireless

5G NR specification is still under development in 3GPP (3GPP TS 38.201-36.215 2017). The technical studies (3G3GPP TR 38.801 2017; 3GPP TR 38.802 2017; 3GPP TR 38.803 2017; 3GPP TR 38.804 2017; Ericsson 2017) completed in March 2017 have summarized very well the agreed NR framework. In this section, we will provide high-level agreement of 5G wireless from access scheme, massive MIMO, and signaling aspects.

Access Scheme

5G NR has a dynamic frame structure with flexible numerology that is scalable for a wide range of spectrums and applications. The key components consist of waveform, frame structure, modulation, and coding schemes.

5G NR waveform is Orthogonal Frequency Division Multiplexing (OFDM) with a cyclic prefix (CP) in both downlink and uplink up to at least 52.6 GHz. The same waveform on downlink and uplink has simplified the overall design, in particular, inclusive of wireless backhauling, device-to-device, and vehicle-to-X applications. Additionally, discrete Fourier transform spread OFDM



5G Wireless, Fig. 2 Flexible and scalable resource allocation

(DFT-s-OFDM) is also supported in uplink to match LTE coverage. Windowing or filtering can be applied on top of OFDM to improve spectrum confinement and enable spectrum utilization higher than 90% of current LTE.

The 5G NR subcarrier spacing, Transmission Time Interval (TTI), and cyclic prefix are flexible and configurable based on deployed spectrum and service requirement, as shown in Fig. 2. The subcarrier spacing can be 15×2^k extended from fixed 15 kHz in 4G LTE, while k is an integer. Thus it supports ultra-wide band for the Gbps enhanced broadband service and fixed wireless communications. It will allow a shorter time slot and larger subcarrier spacing for a delay-sensitive mission critical application with a finer time granularity. It also allows a narrow bandwidth and smaller subcarrier spacing for massive IoT connections to extend coverage and reduce device complexity and power consumption.

The NR frame structure supports a self-contained transmission; thus the reference signals required for data demodulation are included in the given slot or beam. It also supports well-confined transmissions in time and frequency, avoiding control channels mapping across full system bandwidth. It avoids static timing relation across slots and different transmission directions and removes the resource allocation limitation of predefined transmission time.

NR supports the same modulation schemes as that of LTE, which includes QPSK, 16-QAM, 64-QAM, and 256-QAM. It supports $\pi/4$ -BPSK

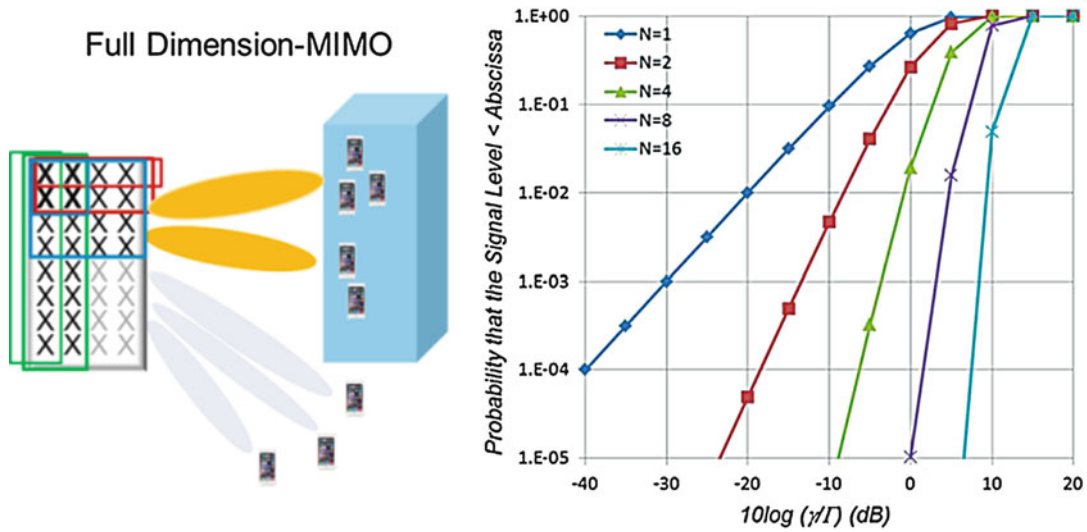
for DFT-s-OFDM for extended coverage. It will define 1024-QAM for fixed wireless applications. The final NR specification will cover a more extensive modulation scheme with different user equipment (UE) categories for diverse use cases.

NR supports low-density parity-check (LDPC) codes for the data channel and polar codes for the control channel. Those have evolved from traditional convolutional and turbo codes used in LTE. The LDPC has showed much superior performance at a wide range of coding rates. It supports high peak rate and low latency at high coding rate and high coding gain with high reliability at low coding rate. Polar codes are the first class of codes shown to achieve the Shannon capacity. It will be used for Physical Broadcast Channel (PBCH) and control information with shorter block length at a reasonable decoding complexity.

Massive MIMO

The massive MIMO provides a means to support high reliable services and achieve the three times spectral efficiency enhancement required by IMT-2020. NR beamforming can be applied to both data transmissions and initial access.

NR will define various antenna solutions depending on spectrum. For sub-6 GHz, most likely up to 32 transmitter chains with FDD spectrum are used. The spectral efficiency is improved by high-resolution channel state information (CSI) reporting and multiuser MIMO (MU-MIMO). For higher frequencies, a larger number of antenna (>128) can be supported in a given aperture. TDD spectrum- and reciprocity-based schemes are assumed. mmWave with an analog beamforming is typically limited to a single beam per slot per radio chain and thus requires beamforming on both transmitter and receiver. To support those diverse user cases across 100 GHz, NR defined a highly flexible and unified CSI framework to reduce the coupling among CSI measurement, CSI reporting, and the actual downlink transmissions. This framework also supports multipoint transmissions and coordination. The self-contained frame structure allows network to seamlessly change the transmission point or beam as devices moving across radio nodes.



5G Wireless, Fig. 3 Massive MIMO and benefit

Full-dimension MIMO (FD-MIMO) can significantly increase capacity and coverage and improve reliability and network resiliency, on both horizontal and vertical domains. As we can see from Fig. 3, the more antenna branches, the less fading margin required to meet a reliability requirement. For example, only 3 dB fading margin is required for 16 antennas at 99.99% reliability, while 20 dB margin required for 2 branches and 40 dB required for single branch to achieve the same reliability. It is almost impossible to achieve 5 nine reliability in a fading channel with single branch antenna, while only require 5 dB margin for a 16-branch receiver.

Massive MIMO also brings challenges as the beamforming algorithms are complicated. The device and eNB complexity does not scale well with the number of increased antenna elements. RF and baseband design for massive MIMO at network, as well as at user equipment, are still at development stages. Mobility and connectivity over very narrow beams require fast tracking and switching. Front-haul bandwidth in tens of Gbps is required for the Gbps data transmission among radio, baseband, and network. Thus a lower-layer splitting with compression is necessary to enable advanced receiver at base station while maintaining reasonable front-haul throughput and latency requirement. The standardized lower-layer split-

ting will enable operator to mix and match radio units and base station processors from different vendors.

Dynamical antenna array configurations with flexible number of beams on vertical and horizontal planes are required to support various deployments as showed in Fig. 3. Simulations in 3GPP have shown that for a dense urban skyscraper with vertical traffic distribution, the vertical dominant antenna ports provided more gain, while for typical suburban low-rise deployment, more horizontal ports will provide high capacity. Those will serve as a stepping stone toward 5G by gradually increasing the number of antenna ports from current 4 to more than 32 transmission ports. This will prepare us for 5G massive number of antenna ports.

Reference Signals and Signaling Efficiency

5G NR has much lean reference signals and signaling overhead that translates directly to a better spectral efficient. The always-on broadcast signals are minimized. Common reference signals are replaced by UE-specific reference signals with improved feedback accuracy and less overhead.

The four main reference signals in NR are demodulation reference signal (DM-RS), phase-tracking reference signal (PTRS), sounding

reference signal (SRS), and channel state information reference signal (CSI-RS). DM-RS is used to estimate the radio channel for demodulation. It is UE-specific, only transmitted on an as-needed basis, for example, the transmit interval could be long for low-speed user, while short for high-speed mobile users. PTRS is introduced to compensate UE-specific oscillator phase noise for mmWave. The SRS on uplink and CSI-RS on downlink are used to conduct CSI measurement for a scheduler and link adaptation. They will also be utilized for massive MIMO precoder design and beam management on uplink and downlink, respectively. It is expected 5G NR will improve spectral efficiency by at least 10% from those UE-specific reference signals.

5G NR also reduces signaling overhead by introducing a new RRC inactive state. Thus it can save control signaling by 15% and decrease signaling latency. It also replaces the always-on broadcast signals with on-demand system information.

This reduced signaling increases the system stability, particularly in an overloaded scenario (Yang et al. 2016). Thus, 5G NR is a more efficient, reliable, and robust communication means.

One Network for a Variety of Services

One operational network is a cost-efficient way to engineer, maintain, and operate radio networks from the perspective of capital and operational expense. The 5G wireless network can be used for eMBB, mMTC, and URLLC applications in a wide area radio access, D2D, V2X, and wireless backhaul use cases.

Wireless networks are evolved toward both centralized resource sharing and distributed edge mobile computing. Artificial intelligence is introduced to enable self-organizing at radio accesses, software-defined networking, and network function virtualization at core networks to allow the network to be more dynamic and intelligent with service and content awareness.

5G wireless will be a combination of evolved LTE network with new radio access technologies targeting for wide bandwidth, expanded spec-

trum, enhanced mobile broadband, as well as mission critical connections. This access network with intelligence at both edge and cloud can serve entire communities and industries, supporting both Gigabit Society and the Internet of Everything.

Key Applications

5G wireless will support varieties of applications from ultra-broadband services, to mission critical applications, as well as massive Internet of Things. It will support above 100 MHz instantaneous bandwidth from hundreds of MHz to GHz radio frequency ranges. 5G will profoundly enrich our daily life.

Cross-References

- [Internet of Things](#)
- [Massive MIMO](#)
- [Millimeter-wave Communications](#)
- [Mobility Management](#)
- [Multiple Access Techniques](#)
- [Quality of Service](#)
- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

References

- 3G3GPP TR 38.801, Study on New Radio (NR) Radio Access; Radio access architecture and interfaces, v.2.0.0, March 2017
- 3GPP TR 38.802, Study on NR; Physical layer, v.2.0.0, Mar 2017
- 3GPP TR 38.803, Study on NR; RF and Co-existing, v.2.0.0, Mar 2017
- 3GPP TR 38.804, Study on NR; Radio interface, v.2.0.0, Mar 2017
- 3GPP TS 38.201-36.215, 3GPP, Technical specification group, Radio access networks; NR; Physical Layers; Release 15, Aug 2017
- Ericsson (2017) 5G new radio, designing for the future. Ericsson Technology Review #7
- Verizon (2017) State of the IoT market – 2017. <http://www.verizon.com>
- Yang, J, Song, L, Koeppe, A (2016) LTE field performance for IoT applications. In: Proceedings of VTC, 18 Sept

Flow Table Exhaustion Attacks

► Flow Table Overflow Attacks

Flow Table Overflow Attacks

Qi Li¹, Jiahao Cao¹, Mingwei Xu¹, and Kun Sun²

¹Tsinghua University, Beijing, China

²George Mason University, Fairfax, VA, USA

Synonyms

Flow table exhaustion attacks; Flow table overload attacks

Definition

Flow table overflow attacks consume flow tables that forward and process packets of flows in software-defined networking (SDN), which results in no space left for other flows to install flow rules and thus incur network denial-of-service (DoS). Such attacks are serious security threats in SDN, as they can be easily launched by an adversary having or compromising hosts in the target network.

Historical Background

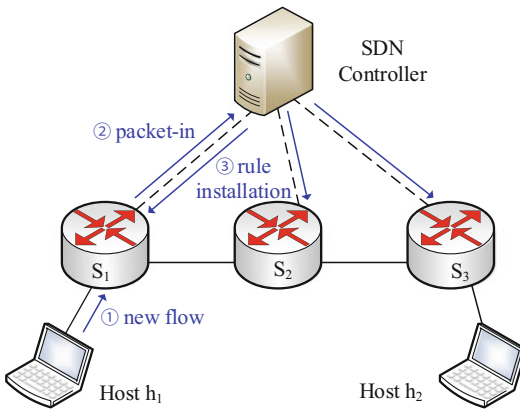
The flow table overflow attack is a kind of table overflow attacks. Table overflow attacks have existed for decades after the Internet was born in 1969. The most common table overflow attack in legacy computer networks is the MAC address table overflow attack. Such an attack is launched by an adversary that sends lots of packets spoofing source MAC addresses to a switch, and thus the switch MAC address table is fully loaded. Another similar attack is the routing table overflow attack. In such an attack, an adversarial

router advertises routes to nonexistent destinations to authorized routers, which can saturate the routing table.

Flow table overflow attacks was newly discovered after the first software-defined networking (SDN) standard, i.e., OpenFlow (McKeown et al. 2008), was announced in 2008. Later on in 2013, studies Kreutz et al. (2013) and Kloti et al. (2013) showed concerns on flow table overflow attacks that can disrupt SDN. In particular, Shin and Gu (2013) designed and evaluated the first feasible flow table overflow attack in SDN, i.e., SDN Scanner, which can fingerprint SDN networks and then generate packets with random header fields to exhaust flow tables. In 2017, Pascoal et al. (2017) improved the attack so that it can be launched with low-rate traffic. At the same time, Cao et al. (2017) provided the LOFT attack, which is sophisticated and can exhaust flow table with a minimal rate in a stealthy way.

Foundations

Flow table attacks are serious security threats in SDN. To understand the attacks, the necessary context should be explained. Hence, a brief introduction to SDN must be firstly noted. SDN is an emerging network paradigm, which decouples the control plane from the data plane. The control plane consists of a centralized controller responsible for taking over the whole network. The data plane contains SDN switches that forward packets. The flow table in SDN switches plays a key role on supporting SDN abstraction, as almost all decisions made by the controller, e.g., how to route flows, are translated into flow rules in the flow table. To support line-speed forwarding and flexible packet processing, dedicated memories called ternary content-addressable memory (TCAM) are used to form the flow table. However, such memories are power-hungry and expensive. Currently, most commercial hardware switches can only store a limited number of flow rules in TCAM (typically, 1500 to 3000 rules (Pascoal et al. 2017)), which has potential risk to be overflowed. Moreover, some network applications in the SDN controller adapt fine-



Flow Table Overflow Attacks, Fig. 1 The proactive rule installation mechanism in SDN

grained control. For example, a firewall application may install flow rules according to IP addresses and TCP ports to block or allow some specific flows. Such kind of application exacerbates the potential risk of flow table overflow.

On the other hand, the proactive rule installation mechanism in SDN can be easily leveraged by an adversary to overflow the flow table. Figure 1 illustrates the mechanism. When a packet from a flow arrives at the ingress switch, it will be forwarded and processed by the related flow rules. If no rule matches the packet, a packet-in message generated by the switch will be reported to the controller for querying flow rules. In normal cases, new flow rules will be directly installed into the switch. However, when the flow table has no free space, the flow rule will be rejected by the switch or installed at the expense of deleting an existing flow rule, which depends on the behavior of switches. Either behavior can degrade the throughput of flows when the flow table is overflowed and thus incurs network DoS. Typically, there are three types of attacks that leverage the above mechanism to overflow flow table:

- *Brute-force flow table overflow.* In such an attack, a compromised host in the target network generates a huge number of anomalous packets. The header fields of each packet are

spoofed as random values. Thus, such packets have high probability of matching no flow rules, which means many new flow rules will be installed into switches. Consequently, the flow table of switches can be quickly saturated.

- *Slow flow table overflow.* The attack is an improvement of the brute-force attack. The brute-force attack needs to generate a huge number of anomalous packets with a high rate. In order to decrease the attack rate, the attack assumes that an adversary controls many hosts in a botnet so that each host can send a unique flow to a common destination. Hence, to route these unique flows, two flow rules (incoming and outgoing rules) for each unique flow will be installed into each switch along with the routing path. To make the attack successfully overflow flow table, it requires the number of the compromised hosts is bigger than half of the number of flow rules that a switch can support.
- *Sophisticated flow table overflow.* The sophisticated attack overflows flow table with a minimal rate in a stealthy way. It contains two phases: probing and attacking. In the probing phase, it generates probing packets to infer the controller's logic of installing flow rules and the survival time of each installed rules. In the attacking phase, a compromised host generates packets to stealthily overflow flow table. Such packets are carefully crafted according to the previously inferred controller's logic of installing rules so that each packet can trigger a new rule installation. Moreover, the minimal packet rate of keeping flow tables overflowed over time are also calculated based on the survival time of flow rules. Thus, packets are sent in a way where flow tables can be slowly overflowed with a minimal attack rate.

Key Applications

Overflowing flow tables can be used as a method to disrupt the target SDN and incur network DoS.

Cross-References

- [Software-Defined Wireless Networking](#)

References

- Cao J, Xu M, Li Q, Sun K, Yang Y, Zheng J (2017) Disrupting SDN via the data plane: a low-rate flow table overflow attack. In: International conference on security and privacy in communication systems. Springer, pp 356–376
- Kloti R, Kotronis V, Smith P (2013) Openflow: a security analysis. In: 2013 21st IEEE international conference on network protocols (ICNP). IEEE, pp 1–6
- Kreutz D, Ramos F, Verissimo P (2013) Towards secure and dependable software-defined networks. In: Proceedings of the second ACM SIGCOMM workshop on hot topics in software defined networking. ACM, pp 55–60
- McKeown N, Anderson T, Balakrishnan H, Parulkar G, Peterson L, Rexford J, Shenker S, Turner J (2008) Openflow: enabling innovation in campus networks. *ACM SIGCOMM Comput Commun Rev* 38(2):69–74
- Pascoal TA, Dantas YG, Fonseca IE, Nigam V (2017) Slow TCAM exhaustion DDOS attack. In: IFIP international conference on ICT systems security and privacy protection. Springer, pp 17–31
- Shin S, Gu G (2013) Attacking software-defined networks: a first feasibility study. In: Proceedings of the second ACM SIGCOMM workshop on hot topics in software defined networking. ACM, pp 165–166

Flow Table Overload Attacks

- [Flow Table Overflow Attacks](#)

Fog

- [Fog Computing for Intelligent Mobile and IoT Networks](#)

Fog Computing

- [Architecture and Data Management for Smart Community Information Platform](#)
- [Data-Driven Edge Computing](#)

Fog Computing for Intelligent Mobile and IoT Networks

Yuan-Yao Shih², Hung-Yu Wei³, and Ai-Chun Pang^{1,2}

¹Department of Computer Science and Information Engineering, Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

²Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan

³Department of Electrical Engineering and Graduate, Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan

Synonyms

[Edge computing](#); [Fog](#); [Fog networks](#)

Definitions

Fog computing is an extension of the cloud computing paradigm from the core of the network to the edge of the network. It is an architecture that utilizes collaborative resource-limited devices (e.g., mobile end-devices, access points, routers, switches) or more capable machines (e.g., cloudlets) at the near-user edge to provide storage, communication, and computing resource for services and applications to users.

Historical Background

In the past years, computing paradigms have evolved from distributed and parallel to cloud computing, which can provide elastic resources to applications on the resource-limited end-devices. However, the connection between the cloud and end-devices over the Internet is not suitable for latency-sensitive applications, such as wearable computing, smart metering, smart home/city, vehicles, and health monitoring. That

is because existing communication systems cannot meet the ultralow latency requirements. For example, LTE can achieve a typical round-trip latency of 25 ms, while many ultralow-latency applications require less than 10 ms of latency. Thus, although, nowadays, cloud computing is widely used, many challenges remain unsolved, such as mobility support, location awareness, and ultralow latency requirements.

Fog computing is then proposed by Cisco (Bonomi et al. 2012) in 2012 to overcome those challenges by migrating the computing from the cloud to the edge of the network. It enables computing and supports service virtualization at the edge of the network, closer to the end-devices of users. On the other hand, in 2014, the work (Chiang 2014) from Princeton University proposes *fog network* as a similar concept of the fog computing while focusing more on utilizing one or many end-user clients or near-user edge devices to handle storage, communications, control, configuration, measurement, and management function of the IoT collaboratively. Meanwhile, *Fog-Radio Access Network (F-RAN)* was introduced at the Next Generation Mobile Network (NGMN) Forum in 2014 (Shih et al. 2017). The F-RAN leverages the equipment in the radio access network (RAN) on small cells to be aggregators connecting the IoT devices and the cellular network. Moreover, many standardization organizations, such as European Telecommunications Standards Institute (ETSI), and research projects, such as a European Union project named “Distributed Computing, Storage, and Radio Resource Allocation over Cooperative Femtocells (TROPIC),” also focus on the idea of “extension of computation to the edge of the network.” To consolidate those efforts on developing fog computing, in 2015, OpenFog Consortium was founded by ARM Holdings, Cisco Systems, Dell, Intel, Microsoft, and Princeton University to promote interests and development in fog computing. OpenFog particularly emphasizes the openness and hierarchy where implementations of the fog computing architecture can reside in multiple layers of a network’s topology.

There are similar concepts such as Mobile Cloud Computing (MCC) (Dinh et al. 2013) and Multi-Access Edge Computing (MEC) (Taleb et al. 2017). MCC moves the data storage and processing from mobile devices to more powerful devices, while MEC focuses on deploying resource-rich fog servers like cloudlets at the edge of mobile networks. MEC was initiated in 2014 by ETSI under the name of Mobile Edge Computing; then its scope has been expanded in 2017 to encompass nonmobile network requirement and replace the Mobile with Multi-Access in the name. On the other hand, fog computing is a more generalized computing paradigm such that the computation is performed at the edge of the network, but it is not defined who exactly performs the computing.

Foundations

Architecture of Fog Computing

A typical architecture of the fog computing composed of the access network and end-devices of users. There are radio access equipments, such as Wi-Fi access points or femtocell base stations and remote radio heads (RRHs) extended from existing macro-eNodeBs, at the access network, serving as the middle layer between the end-device and the cloud. Those equipments can communicate with each other directly via advanced wireless technologies, and they are all connected to the core network and the Internet via the back-bone links. The end-devices of users (such as set-top boxes, network-attached storage, wearable devices, smartphones, or IoT smart devices) are connected to the access points or base stations via wireless technologies. The cloud behind the core network can provide non-real-time computing and big data analysis as the fog is considered as an extension, not a replacement, of the cloud.

To provide Quality-of-Service (QoS) and enable flexible operation, fog computing has many flexible and scalable options for computing resource deployment. The computing resources can be provided by every role in the network. That includes the existing radio access equipments and extra dedicated servers (e.g.,

cloudlet) deployed beside those equipments at the access network. In addition, users can also provide the sparse computing resource of their end-devices. Those devices, which are the basic units to provide computing resource, are denoted as fog computing nodes (FNs). Also, there exists the fog controller, which can be any FNs or independent servers in a centralized or distributed fashion, which is mainly in charge of receiving service requests, distributing tasks to FNs, and summarizing processing results.

A typical IoT application service model of the fog computing is that, first, the IoT device initiates the service request to the fog controller and, then, the fog controller distributes the computing-intensive tasks of the service to one or multiple FNs. The FNs process those tasks and return the outcomes to the fog controller. Finally, the controller summarizes the received outcomes and then sends the final outcomes back to the IoT device.

Collaboration Between Fog and Cloud

Fog computing can be collaborated with cloud computing such that FNs might be integrated with cloud computing infrastructure and IoT devices. Some computing tasks will be carried locally at IoT devices; meanwhile, some computing tasks will be served by FNs, and some might be forwarded to cloud computing infrastructure. We call the system with IoT devices, fog nodes, and cloud computing as a Thing-Fog-Cloud continuum. The computational tasks could be divided and distributed in this Thing-Fog-Cloud continuum. Depending on the computational loading, networking congestion, and application requirements, the computing tasks might be allocated to suitable nodes in this Thing-Fog-Cloud continuum.

Meanwhile, there are diverse IoT applications with various delay requirements served by the common infrastructure. There will be delay-tolerant and delay-sensitive IoT applications coexisting in the wireless fog computing environment. For example, in the fog-enabled factory automation environment, robotic control IoT application is delay sensitive, while collecting manufacturing parameters for IoT process

enhancement is delay tolerant. With the mixed delay-sensitive and delay-tolerant data packets, networking infrastructure needs to be able to differentiate those packets and prioritize their transmission. Besides, computational resource allocation framework should also differentiate the serving policies for delay-tolerant and delay-sensitive tasks. The differentiation and management mechanism might be implemented in IoT device level or in fog nodes. For example, a fire alarm or pressure alarm sensor might be treated differently than a weight sensor that measures the weight for a box of products. In another example, a fog node that receives videos from a surveillance camera might be triggered to transmit a delay-sensitive or a delay-tolerant message, based on the video recognition results from the received surveillance videos. In addition, the delay-sensitive emergency videos might need to be further analyzed with high priority in a fog node; meanwhile, the non-emergency video clips could be forwarded to cloud computing infrastructure to be stored for record or for future analysis. With the flexible management and orchestration in the fog computing system, intelligent IoT applications could be served (Ku et al. 2017).

Enabling Technologies of Fog Computing

Fog computing also relies on some enabling technologies that allow flexibility and multi-tenancy support:

- *Virtualization:* Virtualization in computing infrastructure and networking systems has been observed as an emerging trend. More and more network infrastructure nodes might be implemented with general-purpose hardware and achieve adaptive configurability through software. With this trend, fog computing technologies might become possible by utilizing those general-purpose hardware platforms. Classic virtualization technology, virtual machine (VM), consumes a significant amount of resources and has slow startup process because it needs to boot a complete OS. Different from the traditional VM,

the emerging container technology, whose abstraction takes place at the OS level, is a lightweight virtualization solution that enables portable runtimes of fog services, which is particularly useful for mobile users (Xavier et al. 2013).

- *Network Function Virtualization (NFV)*: NFV is a technology that enables decoupling network functions and services from proprietary hardware (Han et al. 2015). With NFV, hardware node could be used to provide networking functionality in software as well as fog computing services. Also, fog computing can also leverage NFV for managing the virtualized instances related to specific applications and services. For example, increasing the resources for a popular service can be easily achieved by adding another software instance or particular resources, e.g., computing power, storage, or network bandwidth.
- *Software-Defined Networks (SDN)*: SDN is a technology that decouples the control plane from the data plane and enables network programmability through the use of standard APIs (Kreutz et al. 2015). SDN introduces a logically centralized control, abstracting the underlying network infrastructure, and enables real-time network control. With SDN, fog controller can have flexible and efficient management for data routing.

Key Applications

Fog computing paradigm and its integration with wireless networks are aimed to provide low-latency IoT services. By placing computing resources that are close to IoT devices, some delay those fog computing nodes that could serve sensitive computing tasks. The following are some key applications of fog computing:

- *Mobile augmented reality*: Due to the widespread availability, portability, and built-in good-quality cameras of smartphones, mobile augmented reality (mobile AR) has been drawing many attentions. Mobile AR,

enabling virtual graphic imagery to overlay physical object in real time, makes many new applications (e.g., navigation, public security, real-time translation) possible. Some popular products or projects include Google Glass and Microsoft HoloLens that have been introduced. One of the core components for mobile AR is the pose tracking module, in charge of calculating the relative pose of the camera in real time for the location and orientation of the virtual components. Typical CPU hardware on smartphones is not able to carry the computational complexity of tracking algorithm; thus, the fog computing can provide the necessary application-layer computing power via distributing the computing-intensive tasks of the mobile AR to multiple FNs (Pang et al. 2017).

- *Traffic control system*: A smart traffic control system enables traffic signals to adapt to actual traffic demand. The system needs very fast interactions such that traffic lights can respond to the fast-changing traffic conditions and accident-avoidance alerts can be given promptly. Fog computing can provide near-edge computing power to those traffic lights and roadside sensors for the ultralow-latency decision-making (Liu et al. 2018).
- *Content caching*: In fog computing system, caching mechanism is also considered as a useful design for latency reduction, especially for the application for information query and downloading (Tandon and Simeone 2016). One such application is a video surveillance system that recognizes the people entering a building through video matching with the personal database. As people go to their office area daily, caching mechanism in a local fog node will be helpful for pattern matching and basic information query in the nearby area.
- *Autonomous and connected cars*: With the emerging trend of autonomous driving and advancement of vehicular communication technologies, IoV (Internet of Vehicles) becomes an important IoT deployment scenario (Kaiwartya et al. 2016). Many applications for IoV (e.g., augmented reality for information providing through a heads-up

display and accident avoidance in autopilot) not only need the near-real-time response but also require a tremendous amount of computing power to process a significant amount of data generated by the sensors on the vehicles. Thus, fog computing might play a vital role in the IoV deployment via moving the intelligence closer to the vehicles, as many IoV applications are delay sensitive. In the future, some vehicles might be equipped with powerful computing hardware, such as GPU or FPGA, and some vehicles might only have limited computing power. Fog computing can be a suitable solution for enhancing computational capabilities in vehicular environments via moving the intelligence closer to the vehicles. To provide low-latency IoV services, some roadside units might also be capable of serving as a fog computing node, and some cellular base stations might be tightly integrated with fog servers in the radio access network or in the core network. V2X (Vehicle-to-Everything) communications might need to be tightly integrated with fog computing resource in this scenarios.

- *Wearable application and service offloading:* Wearable devices have become ubiquitous and provide convenient features for users, such as heart rate monitor and fitness tracker. However, since those devices are designed to be lightweight and power-saving, the computing capability of wearable devices is often limited. The existing solution is to pair with the local hub, mostly smartphone, via Bluetooth Low Energy (BLE) or Wi-Fi interface. Whenever a request takes place, the wearable device will automatically transfer it to the local hub and then reply with the results. Connection via Bluetooth, however, is only functional within a narrow connection range, whereas connection via Wi-Fi interface has a long response time. It means when someone left the smartphone at home and walked out with the wearable device only, the services may suffer from prolonged network latency delay. Fog computing can be helpful to address those shortcomings via deploying the wearable services at the

edge side to serve the wearable devices (Lin et al. 2016).

- *Tactile Internet:* The Tactile Internet requires ultra-responsive and ultrareliable network connectivity that enables it to deliver physical haptic experiences remotely (Simsek et al. 2016). The concept of Tactile Internet gives us a sneak preview into many future promising applications, such as exoskeletons, remote surgery, virtual sport training, and remote robot controlling.

Future Directions

As fog computing is still in its infant stage, most future needs and challenges will come in the areas of fog computing. Listed below are some future challenges:

- *Joint networking, storage, and computing resource allocation:* Close integration between wireless networking infrastructure and fog computing will reduce the overall end-to-end latency for real-time IoT applications. As a result, joint resource allocation for wireless radio resource and computing resource is critical to meet the low latency requirements for delay-sensitive IoT applications. For example, to minimize the end-to-end application latency, assigning a computing task to the nearest fog node might not be optimal. The optimal computing task assignment mechanism might consider the wireless channel quality to a certain base station or access point, network latency in wired-line backhaul link, and computing load in a fog node. Moreover, when caching mechanism is applied, fog storage needs to be considered in resource allocation and management, in addition to computing and wireless network resource allocation.
- *Platform management and orchestration mechanism:* Between IoT devices and cloud computing servers, there might be possibly multiple tiers of fog nodes. The computing tasks might be allocated through a multi-

tier fog hierarchy or offloading from a fog node to another peering fog node. Thus, in fog computing systems, management and orchestration mechanism is essential for the smooth operation of IoT applications with different requirements that coexist in the fog computing environment.

- *Mobility support:* Compared to factory automation applications, in which devices are mostly static, devices in vehicular IoT applications or wearable applications move with mobility. Those IoT devices might connect to their best-serving fog node at one moment that might become sub-optimal configuration as the vehicle or individual moves. As a result, the fog computing deployment needs to consider the mobility issue. In addition, mobility prediction and service provisioning might be included for better QoS of fog deployment.
- *Security and privacy:* Security and privacy are also important issues in the fog. Many security and privacy issues of the cloud are inherited in fog computing since the fog is considered as an extension of the cloud. Although some issues can be addressed using existing schemes, new security and privacy challenges arise due to the distinct characteristics of fog computing, such as mobility, heterogeneity, large-scale geo-distribution, location awareness, and ultralow latency. For example, how to design access control across client-fog-cloud with restriction is challenging. In addition, the location privacy of the FNs and the users is also a critical issue as task offloading takes place with consideration of their locations for lowering the latency.
- *Business model and incentive design:* Fog computing cannot be prosperous without a sustainable business model. The computing platform consists of the following three parties: (1) the operator, who owns the radio access equipment at the edge of the networks; (2) application service providers, who operate delay-sensitive applications; and (3) users who enjoy the applications. The operator can provide fog computing service to the delay-sensitive applications, run by the application

service providers, at near edge of the users via utilizing the computing resource of those radio access equipments. The business model needs to ensure every party can gain a positive benefit from joining the service. Two main issues of the model need to be resolved: first, the operator needs to know its cost, i.e., the amount of resource needed to fulfill the service requests of an application within a certain degree of latency. Second, the price of the service should be set as both operator and application service provider agree.

Cross-References

- [Big Data in 5G](#)
- [Data-Driven Edge Computing](#)
- [Evolution of Vehicular Networks in the Transition from 3G to 4G Systems](#)
- [Mobile Edge Caching in HetNets](#)
- [Mobile Edge Computing: Low Latency and High Reliability](#)
- [Wireless Edge Caching](#)

References

- Bonomi F, Milito R, Zhu J, Addepalli S (2012) Fog computing and its role in the internet of things. In: Proceedings of ACM SIGCOMM workshop on mobile cloud computing, pp 13–16
- Chiang M (2014) Fog networks. Technical report, Princeton University
- Dinh HT, Lee C, Niyato D, Wang P (2013) A survey of mobile cloud computing: architecture, applications, and approaches. *Wirel Commun Mob Comput* 13(18):1587–1611
- Han B, Gopalakrishnan V, Ji L, Lee S (2015) Network function virtualization: challenges and opportunities for innovations. *IEEE Commun Mag* 53(2):90–97
- Kaiwartya O, Abdullah AH, Cao Y, Altameem A, Prasad M, Lin CT, Liu X (2016) Internet of vehicles: motivation, layered architecture, network model, challenges, and future aspects. *IEEE Access* 4:5356–5373
- Kreutz D, Ramos FMV, Verssimo PE, Rothenberg CE, Azodolmolky S, Uhlig S (2015) Software-defined networking: a comprehensive survey. *Proc IEEE* 103(1):14–76
- Ku YJ, Lin DY, Lee CF, Hsieh PJ, Wei HY, Chou CT, Pang AC (2017) 5g radio access network design with the fog paradigm: confluence of communications and computing. *IEEE Commun Mag* 55(4):46–52

- Lin HP, Shih YY, Pang AC, Lou YY (2016) A virtual local-hub solution with function module sharing for wearable devices. In: Proceedings of the 19th ACM international conference on modeling, analysis and simulation of wireless and mobile systems (MSWiM), pp 278–286
- Liu J, Li J, Zhang L, Dai F, Zhang Y, Meng X, Shen J (2018) Secure intelligent traffic light control using fog computing. *Futur Gener Comput Syst* 78:817–824
- Pang AC, Chung WH, Chiu TC, Zhang J (2017) Latency-driven cooperative task computing in multi-user fog-radio access networks. In: 2017 IEEE 37th international conference on distributed computing systems (ICDCS), pp 615–624
- Shih YY, Chung WH, Pang AC, Chiu TC, Wei HY (2017) Enabling low-latency applications in fog-radio access networks. *IEEE Netw* 31(1):52–58
- Simsek M, Aijaz A, Dohler M, Sachs J, Fettweis G (2016) 5g-enabled tactile internet. *IEEE J Sel Areas Commun* 34(3):460–473
- Taleb T, Samdanis K, Mada B, Flinck H, Dutta S, Sabella D (2017) On multi-access edge computing: a survey of the emerging 5g network edge cloud architecture and orchestration. *IEEE Commun Surv Tutor* 19(3):1657–1681
- Tandon R, Simeone O (2016) Cloud-aided wireless networks with edge caching: fundamental latency trade-offs in fog radio access networks. In: 2016 IEEE international symposium on information theory (ISIT), pp 2029–2033
- Xavier MG, Neves MV, Rossi FD, Ferreto TC, Lange T, Rose CAFD (2013) Performance evaluation of container-based virtualization for high performance computing environments. In: 2013 21st euromicro international conference on parallel, distributed, and network-based processing, pp 233–240

Fog Forensics

- [Internet of Things \(IoT\) Forensic Science](#)

Fog Networks

- [Fog Computing for Intelligent Mobile and IoT Networks](#)

4G/LTE Technologies or 4G/LTE Wireless Networks

- [Security and Privacy in 4G/LTE Network](#)

4G: Fourth Generation

- [Security and Privacy in 4G/LTE Network](#)

Free Basics

- [Game Theoretic Modeling of Zero-Rating Internet Provisioning](#)

Frequency Regulation

- [Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future](#)

Frequency-Domain Equalization in OFDM and Single-Carrier Modulation

Rui Dinis¹ and João Guerreiro^{2,3}

¹Department of Electrical Engineering, Instituto de Telecomunicações and FCT-UNL, Caparica, Portugal

²Instituto de Telecomunicações, Lisbon, Portugal

³Universidade Autónoma de Lisboa, Lisbon, Portugal

Synonyms

[Channel equalization for block transmission schemes](#)

Definition

Frequency-domain equalization (FDE) is a technique to mitigate the intersymbol interference (ISI) associated with the channel

effects of the transmission medium, where the equalization is done through frequency-domain processing. The FDE is done digitally by using discrete-time Fourier transforms (DTFTs), which can be efficiently implemented by the fast Fourier transform (FFT) algorithm. For multicarrier (MC) modulations such as orthogonal frequency division multiplexing (OFDM), the FDE procedure occurs naturally, since the data are transmitted in parallel in the frequency-domain. However, FDEs can also be employed with single-carrier (SC) modulations and have been shown to be a great alternative to conventional, time-domain equalizers (TDEs). For this reason, single-carrier with frequency-domain equalization (SC-FDE) and OFDM are nowadays the most popular broadband transmission techniques.

Historical Background

The bit rates and bandwidth of wireless communication systems increased substantially over the last decades, making them an effective alternative to wired communications and allowing users to experience data rates on the order of many MBit/s or even GBit/s. The evolution of wireless communications is intimately related with the improvement of the transmission techniques that support them, which have become more resilient to the multipath propagation, as well as easily implemented in the digital domain through modern signal processing algorithms.

The transmission over frequency-selective channels such as the ones associated with multipath wireless scenarios leads to ISI and, consequently, performance degradation, leading to the need to employ equalizers. The optimum receiver under ISI conditions is the maximum likelihood sequence estimation (MLSE) receiver, also known as Viterbi equalizer, whose complexity grows exponentially with the channel impulse response length. Therefore, MSLE is only suitable for short impulse responses, since its complexity becomes prohibitively high for large constellations and/or channels

with long impulse response (Forney 1972). The conventional approach is to employ TDE schemes (Qureshi 1985), which although simpler than MSLE can still be too complex since they require highly complex filters, composed by a very large number of taps. Therefore, TDEs are not suitable for the highly frequency-selective scenarios that one might expect in broadband wireless systems.

Multicarrier transmission techniques, which were proposed in the mid-1960s (Chang 1966; Saltzberg 1967), but whose importance only started in the mid-1990s, with the strong evolution of electronics and digital signal processing techniques (DSP), are one of the solutions for wireless broadband communications. Among those techniques, the most popular is OFDM (Cimini 1985; Bingham 1990), which is a prefix-assisted block transmission scheme. The key advantage of OFDM schemes is the decomposition of a frequency-selective channel in a large number of parallel channels that behave as flat channels at the subcarrier level, simplifying the equalization procedure.

However, since OFDM signals are formed by a large number of modulated subcarriers, their envelope has substantial peak-to-average power ratio (PAPR), which constitutes a severe problem that can promote the unwanted effects of some radio frequency (RF) impairments, such as nonlinear distortion. In that context, SC-FDE modulation (Sari et al. 1994; Falconer 2002) was proposed as an alternative to OFDM, which can achieve similar performance without the problems associated with nonlinear distortion such as in-band radiation and out-of-band (OOB) radiation.

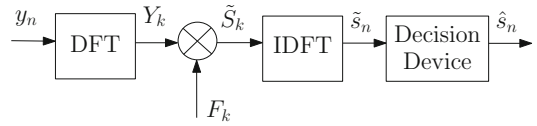
In the last decade, SC-FDE techniques have evolved, allowing performances close to the matched filter bound (MFB). The main enabler of such excellent performance is the concept of iterative block decision feedback equalization (IB-DFE) (Benvenuto and Tomasin 2002), where the equalization is done iteratively to further reduce the ISI levels and improve the performance.

Foundations

The impact of multipath channel on the transmitted signals can be modeled as a linear convolution. However, if the data is transmitted in blocks and a cyclic prefix (CP) longer than the overall channel impulse response (including transmit and receive filters) is appended to each block, then the linear channel convolution behaves as a cyclic convolution with respect to the useful part of the data block. In fact, by adding the CP, which essentially is a repetition of the last part of the data block, we have two main effects: (i) we are able to remove the interference between different blocks, and (ii) the useful part of the received signal (i.e., after removing the CP) is the useful part of the transmitted block times a circulant matrix. Since a circulant matrix can be diagonalized by a discrete Fourier transform (DFT) matrix, in practice this means that there will be no interference at the subcarrier level (or intercarrier interference (ICI)). Therefore, in the frequency-domain (i.e., after a DFT operation on the useful part of the received block), the output of a multipath channel can be seen as the subcarrier-wise multiplication of the channel frequency-response and the transmitted data. The FDE input for a given subcarrier k is given by

$$Y_k = H_k S_k + N_k, \quad (1)$$

where H_k , S_k , and N_k are the channel frequency-response, the data, and noise (assumed additive white Gaussian noise (AWGN)) components associated with the k th subcarrier, simplifying tremendously the channel equalization process, since the channel behaves as flat at the subcarrier level. Clearly, the use of a CP leads to lower energy and spectral efficiencies. However, the benefits of CP in terms of equalization simplicity and achievable performance are considerably higher. For single-carrier modulations, an inverse discrete Fourier transform (IDFT) brings FDE output (i.e., the equalized signal $\tilde{S}_k$) back to the time domain, where the detection is performed and an estimate of the transmitted symbol, $\hat{s}_n$, is obtained. A conventional FDE scenario is illus-



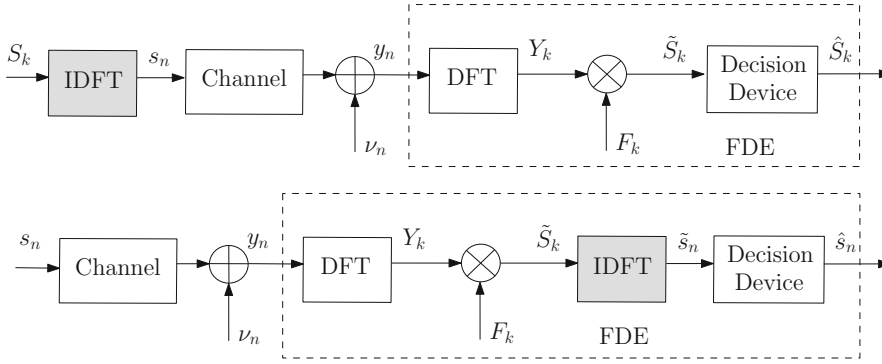
Frequency-Domain Equalization in OFDM and Single-Carrier Modulation, Fig. 1 Conventional FDE scheme

trated in Fig. 1. This FDE forms the essence of SC-FDE schemes, where a data block with an appropriate CP is transmitted and an FDE is adopted at the reception.

However, the FDE is not restricted to SC-FDE and is also at the core of the popular OFDM schemes. In fact, both OFDM and SC-FDE are CP-assisted block transmission techniques with an FDE at the receiver. Moreover, the overall implementation complexity of the two techniques is similar, since both have a pair of fast Fourier transforms (FFTs) (which implement the DFT/IDFT) in their communication chain. Their main difference relies on the fact that, in SC-FDE, both FFTs are placed at the receiver, while in OFDM, since the data is transmitted in the frequency-domain, there is no need to have an IDFT after the FDE. This explains why the reception is more complicated in SC-FDE than in OFDM and why SC-FDE signals can have much lower PAPR than OFDM signals. For this reason, SC-FDE is usually employed in uplink, while OFDM is more adequate for downlink communications, where the energy efficiency requirements are lower. The communication chains of SC-FDE and OFDM are shown in Fig. 2. After the FDE, the equalized signal is sent to the detector that outputs the estimated data signal $\hat{S}_k$ or $\hat{s}_n$ when OFDM or SC-FDE is considered, respectively.

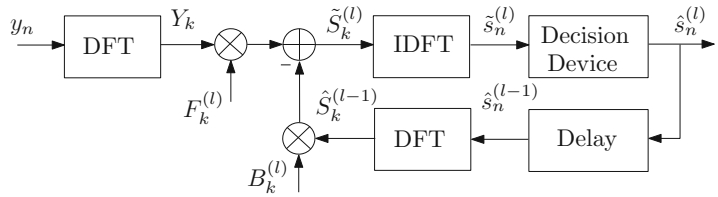
The basic FDE can be regarded as a linear equalizer operating at the subcarrier level. This is suitable for OFDM schemes, where the data symbols are directly transmitted by the subcarriers.

For SC-FDE schemes, the FDE is usually optimized under the minimum mean squared error (MMSE) criteria (Rugini et al. 2005; Pincaldi et al. 2008), which means that there will be some



Frequency-Domain Equalization in OFDM and Single-Carrier Modulation, Fig. 2 Simplified communication chains of OFDM (top figure) and SC-FDE (bottom figure)

Frequency-Domain Equalization in OFDM and Single-Carrier Modulation, Fig. 3 Conventional nonlinear FDE scheme



residual ISI. It is well-known that for SC modulations, nonlinear equalizers outperform linear equalizers, the same applying to SC-FDE. Therefore, it is desirable to have nonlinear FDEs, and several approaches have been proposed to achieve this. The most popular nonlinear FDE is the IB-DFE (Benvenuto et al. 2010), where we have a feedforward part characterized by the coefficients F_k (as with linear FDEs) and feedback part characterized by the coefficients B_k . The general scheme of nonlinear FDEs is represented in Fig. 3. An MMSE-based IB-DFE is characterized by

$$B_k^{(l)} = F_k^{(l)} H_k - 1. \quad (3)$$

Therefore, for IB-DFE techniques, the equalized signal associated with a given subcarrier is given by

$$S_k^{(l)} = F_k^{(l)} Y_k - B_k^{(l)} \hat{S}_k^{(l-1)}. \quad (4)$$

As we increase the number of iterations of the IB-DFE, the residual ISI levels at the FDE output are reduced, which allows performance improvements as we increase the number of iterations, at the expense of some additional and usually affordable computational complexity.

$$F_k^{(l)} = \frac{H_k^*}{(1 - |\rho^{(l)}|^2) |H_k|^2 + \frac{1}{\text{SNR}}}, \quad (2)$$

where $\rho^{(l)}$ is the reliability of the data block detected in the l th iteration, which can be computed as described in Silva et al. (2012), and SNR is the channel's signal-to-noise ratio. The feedback equalization factor is given by

Key Applications

The FDE concept can be used in OFDM and SC-FDE, as well as other block transmission techniques like multicarrier code division multiple access (MC-CDMA) and direct sequence code division multiple access (DS-CDMA), both in

single-antenna and multi-antenna scenarios like multi-input, multi-output (MIMO) schemes.

OFDM and SC-FDE are employed in the physical layer of several systems such as WiMAX (WiM 2004), 4G cellular systems (LTE 2006), and digital television broadcasting (DVB 2009). Moreover, they are the main contenders for the radio access of several 5G use cases.

References

- Benvenuto N, Tomasin S (2002) Block iterative DFE for single carrier modulation. *Electron Lett* 38(19):1144–1145
- Benvenuto N, Dinis R, Falconer D, Tomasin S (2010) Single carrier modulation with nonlinear frequency domain equalization: an idea whose time has come again. *Proc IEEE* 98(1):69–96
- Bingham J (1990) Multicarrier modulation for data transmission: an idea whose time has come. *IEEE Commun Mag* 28(5):5–14
- Chang R (1966) Synthesis of band-limited orthogonal signals for multichannel data transmission. *Bell Syst Tech J* 45(10):1775–1797
- Cimini L (1985) Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing. *IEEE Trans Commun* 33(7):665–675
- DVB (2009) Digital video broadcasting (DVB); framing structure, channel coding and modulation for digital terrestrial television, ETSI EN 300 744 V1.6.1
- Falconer D (2002) Frequency domain equalization for single-carrier broadband wireless systems. *IEEE Commun Mag* 40(4):58–66
- Forney G (1972) Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. *IEEE Trans Inf Theory* 18(3):363–378
- LTE (2006) 3rd generation partnership project: technical specification group radio access network; physical layers aspects for evolved UTRA, 3GPP TR 25.814 V7.1.0.
- Pancaldi F, Vitetta G, Kalbasi R, Al-Dhahir N, Uysal M, Mheidat H (2008) Single-carrier frequency domain equalization. *IEEE Signal Process Mag* 25(5):1349–1387
- Qureshi S (1985) Adaptive equalization. *Proc IEEE* 73(9):1349–1387
- Rugini L, Banelli P, Leus G (2005) Simple equalization of time-varying channels for OFDM. *IEEE Commun Lett* 9(9):1089–1098
- Saltzberg B (1967) Performance of an efficient parallel data transmission system. *IEEE Trans Commun* 15(6):805–811
- Sari H, Karam G, Jeanclaude I (1994) Frequency-domain equalization of mobile radio and terrestrial broadcast channels. In: *Proceedings of IEEE GLOBECOM'1994* (Fall), pp 1–5. San Francisco, CA
- Silva J, Dinis R, Souto N, Montezuma P (2012) Single-carrier frequency domain equalisation with hierarchical constellations: an efficient transmission technique for broadcast and multicast systems. *IET Commun* 6(13):2065–2073
- WiM (2004) IEEE standard for local and metropolitan area networks – part 16: air interface for fixed broadband wireless access systems, IEEE STD 802.16-2004

Full-Dimension MIMO

► Massive MIMO

Fundamental Properties of Wireless Relays and Their Channel Estimation

Rui Wang² and Ping Wang^{1,2}

¹School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore, Singapore

²Department of Information and Communication Engineering, Tongji University, Shanghai, China

Synonyms

One-way relay; Two-way relay; Wireless relay channel; Wireless relay system

Definition

A wireless relay system involves at least three nodes: a source node, a relay node, and a destination node. The relay node assists the transmission of information from the source to the destination. The relay channels include all channels between these nodes.

Historical Background

Due to the complicated transmission media, fading has been considered as a key characteris-

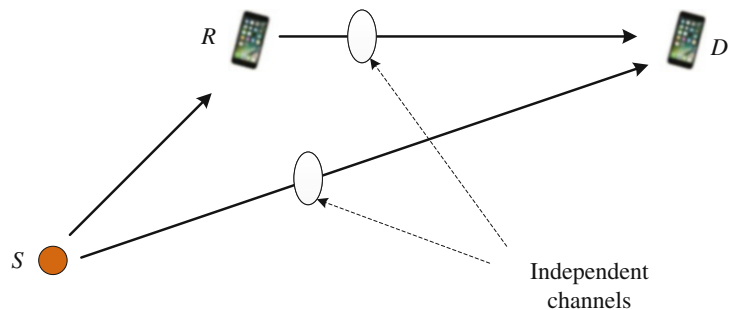
tic feature of the wireless communications. As the wireless fading greatly shortens the delivery distance, the main task in wireless communication systems is to combat the fading to achieve reliable communication. To address this issue, a number of transmission techniques have been proposed. Among them, wireless (radio frequency) relays are widely known to be useful for combating radio channel fading over long distances. Relay technique is a promising technique since it allows a highly flexible equipment deployment to solve the fading problem (Soldani and Dixit 2008). This advantage further enables a cost-effective enhancement of coverage, throughput and network capacity. The relay technique is very helpful in some harsh environments, for example, in rural areas where the population is sparsely distributed and traffic density is low that makes building traditional cellular access networks not economical. There has been a long history in the development of both theory and practice of wireless relays. The earliest concept of wireless relay can be traced back to year 1997 when Scott Guthery used it to construct the relay star network where the relay helps to connect fixed wireless nodes to central server (Guthery 1997). Many of the recent developments and their roots can be traced from Hua et al. (Hua et al. 2012, 2013). Another new angle of understanding the relay assisted communications is from the cooperative communication perspective (Nosratinia et al. 2004; Hong et al. 2007; Sheng et al. 2011). As shown in Fig. 1 where a source node S intends to transmit messages to the destination node D , the source node S sends the signal in the first time slot and the signal is received by both the destination node D and the relay node

R ; then in the second time slot, the relay node forwards the received signal to the destination node D . With an assistance of the relay node R , two copies of signal transmitted from the source S are received at the destination node D where one from S to D and another from S to R then to D . As the channels in those two links are generally independent, two copies of signals can be combined at the destination node D to achieve a more reliable detection. From the communication theory point of view, the performance gain comes from an inherent spatial diversity. As here the wireless relay helps to forward one-way message delivery, i.e., from source to destination, we refer to this channel model as one-way relay channel.

As illustrated in Fig. 1, although one-way relay channel can achieve a more reliable transmission, it requires one more time slot to complete message delivery compared to point-to-point transmission, which sacrifices spectrum efficiency. Later on, to compensate the spectral efficiency loss, in a scenario of two-way transmission where two nodes intend to exchange the messages with each other via a middle-placed relay node, the technique of physical-layer network coding (PLNC) is utilized to improve the spectrum efficiency (Wang and Tao 2012). The relay assisted two-way transmission is referred to as two-way relay channel. As shown in Fig. 2 where source node S_1 desires to transmit signal x_1 to source node S_2 and source node S_2 intends to transmit signal x_2 to source node S_1 , the whole transmission is completed in two time slots. In the first time slot, both source nodes transmit their signals to the relay node simultaneously. In this case, the relay node receives the combination

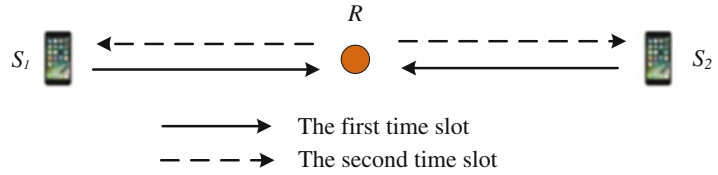
Fundamental Properties of Wireless Relays and Their Channel Estimation, Fig. 1

Wireless one-way relay channel.



Fundamental Properties of Wireless Relays and Their Channel Estimation, Fig. 2

Wireless two-way relay channel



of signals x_1 and x_2 . After performing certain processing, the combination is forwarded to nodes S_1 and S_2 in the second time slot. As two source nodes know their transmitted signals, the self-interference can be canceled from the received combination before decoding the desired signals. It is easy to observe that in two-way relay channel, two messages are delivered within two time slots. In contrast to the one-way relay channel, where two time slots are required to complete one message delivery, two-way relay channel significantly improves the spectrum efficiency.

Several relay processing strategies have been proposed in wireless relay channel. Basically, different strategies usually require different complexity and possess different performance. Two main strategies are amplify-and-forward (AF) strategy and decode-and-forward (DF) strategy. In AF strategy, the relay node simply amplifies the received signal and forwards it directly without decoding the messages. In DF strategy, the relay node decodes the messages from the received signals and regenerates new signals which are sent to the destination subsequently.

As the transmission protocol of the relay channel is quite different from the traditional point-to-point transmission, the corresponding physical layer techniques are greatly modified, especially the channel estimation which is used to obtain the channel state information required for physical layer designs including power allocation (Chen et al. 2017; Ma et al. 2014), precoding design (Cirik et al. 2014; Xu and Hua 2011; Yu and Hua 2010; Rong and Hua 2009; Rong et al. 2009), etc. In wireless relay channel, we generally need to estimate the channels of two-hop transmissions, i.e., from the source to the relay and from the relay to the destination. Lots of new and challenging problems are introduced. In this article, we aim to

provide a brief review of the channel estimation in wireless relay channel.

Foundations

According to transmission protocol, the two-hop channel estimation of wireless relay channel can be performed using the training signals received at the relay and the destination. If the relay node can perform the channel estimation and can transmit the training sequences, the two-hop channel estimation can be decoupled. For example, in the one-way relay channel, the first hop channel, i.e., the channel from the source S to the relay R , can be estimated at the relay node and the second hop channel, i.e., the channel from the relay R to the destination D , can be estimated at the destination node D separately. In this case, the overall channel estimation problem simply reduces to the two point-to-point channel estimations. In the two-way relay channel, we cannot directly decouple the two-hop channel estimation into two point-to-point channel estimation problems as it involves four independent channels. In this case, in the first hop, the channel can be considered as a multiple access channel (MAC) where received training signal at the relay node is used to estimate the channels from the source S_1 to the relay node R and from the source S_2 to the relay node R . While in the second hop, the channel can be treated as a broadcasting (BC) channel where the single training sequence sending from the relay node is received at the destination nodes and is then utilized to estimate the channels from the relay node R to two destination nodes S_1 and S_2 .

In another scenario, if the relay node cannot perform channel estimation or send training sequence, the channel estimation must be conducted at the destination nodes and corresponding channel estimation problem

becomes relatively more complicated. Under this setup, besides estimating the individual channels of two-hop transmission, estimating the combined channels, i.e., the cascade of two-hop channels, is an efficient way to simplify the channel estimation problems. The following brief review of the channel estimation in wireless relay channel is given from four classifications: single-antenna single-carrier case, single-antenna multi-carrier case, multi-antenna single-carrier case, and multi-antenna multi-carrier case.

Single-Antenna Single-Carrier Case

In Gao et al. (2008), the authors considered the channel estimation of single-antenna single-carrier one-way relay channel with multiple relay nodes. Assuming that h_i and g_i denotes the channel from source node S to the i -th relay node D_i and the channel from the i -th relay node D_i to the destination node D , respectively. Instead of estimating h_i and g_i individually, the authors considered to estimate the combined channel $h_i g_i$. The authors studied two estimation criterions, least square (L-S) and minimum mean-square-error (MMSE), by separately treating the unknown channels as deterministic variables and statistical random variables. The estimations of combined channel vector $\mathbf{w}$ for LS and MMSE using the received signal at the destination node $\mathbf{d}_2$ can be expressed as

$$\hat{\mathbf{w}} = \mathbf{A}_{\text{LS}} \mathbf{d}_2, \quad \hat{\mathbf{w}} = \mathbf{A}_{\text{MMSE}} \mathbf{d}_2 \quad (1)$$

where matrices $\mathbf{A}_{\text{LS}}$ and $\mathbf{A}_{\text{MMSE}}$ are related to the training signal and the beamforming matrix used at each relay node; furthermore, matrix $\mathbf{A}_{\text{MMSE}}$ is also related to the statistical information of the noise and channel. According to estimation theory, the authors derived the estimation error covariance matrix of two criterions. The training design was further investigated with an aim to minimize the estimation error, i.e., the trace of the estimation error covariance matrix. Different from the point-to-point channel estimation, the training design in this work included source training sequence design and relay beamforming matrices design. From the problem formulation,

it is found that the optimization of source training sequence design and relay beamforming matrices design can be combined together, which implies that only one matrix variable $\mathbf{C}$ needs to be equivalently optimized in the training design problem. For the LS estimation, the authors proved that the columns of $\mathbf{C}$ should be orthogonal with each other at the optimal solution, and optimal training design can be found. For the MMSE estimation, the authors transformed the training design problem into a convex semi-definite programming (SDP) problem and the optimal solution can be efficiently solved by modern interior point method based on existing convex optimization software.

Later on, the channel estimations of single-antenna single-carrier two-way relay channel were studied in Gao et al. (2009b) and Xie et al. (2014). In Gao et al. (2009b), the authors studied the two-way relay channel taking a time reciprocal property, which means that the channel from the source to the relay is equal to the channel from the relay to this source. With this assumption, the involved four unknown channels reduce to two unknown channels. Denote the channel from the source S_i to the relay node by h_i , similarly to the one-way relay channel estimation in Gao et al. (2008), the authors proposed to estimate the variables $|h_i|^2$ and $|h_1 h_2|$ at terminal S_i instead of estimating h_1 and h_2 . Two channel estimation criterions, i.e., the maximum-likelihood (ML) and linear maximum average effective signal-to-noise ratio (LMSNR), were utilized to obtain the estimations. In specific, the ML estimation treats the unknown channels as two deterministic values and channel estimation is obtained by solving the following problem (take source S_1 as an example)

$$\max_{|h_1|^2, |h_1 h_2|} \log p(z_1 | h_1, h_2) \quad (2)$$

where z_1 is the received signal at source S_1 and $p(z_1 | h_1, h_2)$ is the probability density function (PDF) of z_1 at given h_1 and h_2 . The LMSNR estimation treats the channels as unknown random values with known statistical information, and the channels are estimated by

maximizing the effective SNR. Moreover, the training sequences at two sources were designed aiming to minimize the Cramer-Rao lower bound (CRLB) and maximize the average effective SNR. The results of optimizations showed that the optimal training sequences at two sources should be orthogonal with each other. In Xie et al. (2014), the authors proposed to use the Bayesian approach to estimate the cascaded source-relay-destination channel. The maximum a posteriori (MAP)-based estimation schemes were developed to estimate the cascaded channel and the amplitude of individual source-relay channels with apriori knowledge of channel distribution information. Additionally, to deal with some practical constraints, an iterative least square-MAP algorithm was developed when noise variance was unknown.

Single-Antenna Multi-Carrier Case

The channel estimation of the single-antenna one-way relay channel was studied in Gao et al. (2011) and Wang et al. (2016) with the orthogonal frequency division multiplexing (OFDM) modulation. In this channel setup, the overall end-to-end channel from the source to the destination is the convolution of the channel $\mathbf{h}$ from the source node to the relay node and the channel $\mathbf{g}$ from the relay node to the destination node in the time domain. In Gao et al. (2011), to allow the destination node to estimate $\mathbf{h}$ and $\mathbf{g}$, besides sending the training signals at the source node, the relay superimposes its own training signal to the received training signal and sends them to the destination node. In this case, the transmit signal $\mathbf{r}_t$ at the relay can be represented as

$$\mathbf{r}_t = \alpha \mathbf{r}_r + \mathbf{t}_r \quad (3)$$

where $\mathbf{r}_r$ is the received signal at the relay node, α is the scaling factor used at the relay to satisfy the power constraint, and $\mathbf{t}_r$ is the new training signal superimposed at the relay node. It was shown that the closed-form expressions of estimated channel are generally not available in OFDM setup. To obtain an accurate estimation, an iterative channel

estimation was proposed where $\mathbf{h}$ and $\mathbf{g}$ were separately updated during the iteration. Different from Gao et al. (2011), the channel was estimated jointly with the unknown carrier frequency offset (CFO) and phase noise (PN) in Wang et al. (2016) under the OFDM framework. To enable the joint estimation of MIMO individual channel, CFO and PN at the destination, the training sequences are assumed to be transmitted at both the source and the relay. Let $\mathbf{y}^{[r]}$ denote the received signal at the destination when relay transmits the training signal and $\mathbf{y}^{[s]}$ denote the received signal at the destination when source transmits the training signal. According to the MAP criterion, the estimations of individual channel, CFO, and PN can be obtained by solving the following optimization problem

$$\left\{ \hat{\phi}^{[s-d]}, \hat{\eta}^{[s-d]}, \hat{\mathbf{h}}, \hat{\mathbf{g}} \right\} \propto \arg \min_{\phi^{[s-d]}, \eta^{[s-d]}, \mathbf{h}, \mathbf{g}} p \left(\phi^{[s-d]}, \eta^{[s-d]}, \mathbf{h}, \mathbf{g} | \mathbf{y}^{[s]}, \mathbf{y}^{[r]} \right) \quad (4)$$

where $\phi^{[s-d]}$ and $\eta^{[s-d]}$ denote the combined CFO from source to destination and the combined PN from source to destination; $\left\{ \hat{\phi}^{[s-d]}, \hat{\eta}^{[s-d]}, \hat{\mathbf{h}}, \hat{\mathbf{g}} \right\}$ denotes the estimated version of $\left\{ \phi^{[s-d]}, \eta^{[s-d]}, \mathbf{h}, \mathbf{g} \right\}$; $p \left(\phi^{[s-d]}, \eta^{[s-d]}, \mathbf{h}, \mathbf{g} | \mathbf{y}^{[s]}, \mathbf{y}^{[r]} \right)$ denotes the posterior distribution of the parameters of interests given $\mathbf{y}^{[s]}$ and $\mathbf{y}^{[r]}$. The ambiguities among the estimated PN, CFO, and channels were analyzed and showed that the estimated source-relay channel suffers from a phase ambiguity. Based on this analysis, a hybrid Cramer-Rao lower bound (HCRLB) for analyzing the performance was derived, which can effectively avoid the estimation ambiguities.

Regarding the single-antenna and multi-carrier two-way relay channel, the channel estimation was studied in the framework of OFDM modulation in Gao et al. (2009a). Different from Gao et al. (2011), the authors assumed that the training sequences were only transmitted from two sources while no training sequence was superposed at the relay node. Two estimation schemes, i.e., block-based and pilot-tone-based, were proposed to estimate the

cascaded source-relay-destination channel and the individual channels, respectively. The block-based estimation scheme uses all carriers in one or more OFDM blocks for channel estimation and generally applies to a scenario where the training sequence is long enough. The pilot-tone estimation scheme uses several pilots residing in one OFDM block to estimate the channel and applies to the scenario with a length-limited training sequence. The estimation ambiguities of two schemes were further analyzed. In specific, the authors showed that when the length of training sequence is larger than a threshold, only the sign ambiguity can be introduced and it does not affect the finally data decoding.

Multi-Antenna Single-Carrier Case

When considering multiple antennas at each node, the channel estimation of single-carrier one-way relay channel was investigated in Rong et al. (2012) and Kong and Hua (2011). The challenge of estimating the multi-input multi-output (MIMO) channels lies in the fact that the estimation variables become unknown matrices, while not the unknown values as in the single-antenna case. In Rong et al. (2012), the MIMO channels in source-relay-destination link and the MIMO channel in direct link were estimated without knowledge of the channel statistical information. In particular, the MIMO channel in direct link was estimated using LS criterion. Regarding the source-relay-destination link, according to the parallel factor (PARAFAC) analysis, the bilinear alternating least-squares (BALS) algorithm was proposed to obtain individual MIMO channels for source-relay link and relay-destination link. It was shown that with a mild length of training sequence, the MIMO channel matrices of two hops can be estimated up to permutation and scaling ambiguities. Moreover, the authors proposed to exploit the knowledge of the relay factors to remove the permutation ambiguity. In Kong and Hua (2011), the authors assumed that the statistical channel information was known in prior, and then the linear MMSE (LMMSE) estimation method was proposed to estimate the MIMO channels. To estimate the individual

MIMO channels in each hop of the source-relay-destination link, the authors proposed a two-step estimation strategy where in the first step, the MIMO relay-destination channel is estimated assuming that the relay node is able to transmit training sequences. With the estimated MIMO relay-destination channel, the MIMO source-relay channel is then estimated at the destination node utilizing the training sequence sent from the source. For the first step, the optimal structure of relay training sequence matrix was derived according to the statistical information of the relay-destination channel. While for the second step, an algorithm was developed to compute the optimal training sequence matrix used at the source and the optimal precoding matrix used at the relay.

Later on, the channel estimation of the MIMO single-carrier two-way relay channel was studied in Wang et al. (2015). Similar to Kong and Hua (2011), the authors assumed that the statistical channel information was known in prior under a Kronecker-correlation model. Additionally, the MIMO channels were estimated in a colored noise environment by considering the impact of the antenna correlation and the interference from neighboring users. To estimate each individual MIMO channel, the authors proposed to decompose the bidirectional transmission of two-way relay channel into two phases, i.e., MAC phase and the BC phase. The optimal LMMSE estimators were derived for each phase. Two iterative training design algorithms were further proposed to obtain the training sequences for the general conditions and they were verified to produce training sequences achieving near optimal channel estimation performance. For certain specific practical scenarios where the covariance matrices of the channel or disturbances are of particular structures, the optimal training sequence design guidelines were provided. To assess the estimation performance, the relationship between the estimation performance and the length of training sequences were established, which showed that when the training sequence length is shorter than the threshold, a lower bound of estimation performance exists no matter how to increase the powers.

Multi-Antenna Multi-Carrier Case

The MIMO channel estimation was extended to multi-carrier case for two-way relay channel in Kang et al. (2017). Instead of estimating the individual channels, the authors in this study proposed to estimate the convolution of two MIMO individual channels using the self-interfering link and information-bearing link under the LMMSE criterion. The training sequences were optimized with an aim to minimize the total MSE under the power constraints at the sources and the relay. To obtain the optimal training sequences, the authors derived optimal structure which then converted the training design optimization problem into a tractable convex form.

Key Applications

Wireless relay is one of the fundamental techniques in cellular wireless communications. It can be efficiently used to extend the wireless coverage and enhance the network throughput in harsh environments with low economy cost.

Cross-References

- Cooperative Communications
- Massive MIMO

References

- Chen L, Meng W, Hua Y (2017) Optimal power allocation for a full-duplex multicarrier decode-forward relay system with or without direct link. *Signal Process* 137:177–191
- Cirik AC, Rong Y, Ma Y, Hua Y (2014) On MAC-BC duality of multihop MIMO relay channel with imperfect channel knowledge. *IEEE Trans Wirel Commun* 13(10):5839–5854
- Gao F, Cui T, Nallanathan A (2008) On channel estimation and optimal training design for amplify and forward relay networks. *IEEE Trans Wirel Commun* 7(5):1907–1916
- Gao F, Zhang R, Liang YC (2009a) Channel estimation for OFDM modulated two-way relay networks. *IEEE Trans Signal Process* 57(11):4443–4455
- Gao F, Zhang R, Liang YC (2009b) Optimal channel estimation and training design for two-way relay networks. *IEEE Trans Commun* 57(10):3024–3033
- Gao F, Jiang B, Gao X, Zhang XD (2011) Superimposed training based channel estimation for OFDM modulated amplify-and-forward relay networks. *IEEE Trans Commun* 59(7):2029–2039
- Guthery S (1997) Wireless relay networks. *IEEE Netw* 11(6):46–51
- Hong YW, Huang WJ, Chiu FH, Kuo CCJ (2007) Cooperative communications in resource-constrained wireless networks. *IEEE Signal Process Mag* 24(3):47–57
- Hua Y, Bliss DW, Gazor S, Rong Y, Sung Y (2012) Guest editorial: theories and methods for advanced wireless relays; issue i. *IEEE J Sel Areas Commun* 30(8):1297–1303. <https://doi.org/10.1109/JSAC.2012.120901>
- Hua Y, Bliss DW, Gazor S, Rong Y, Sung Y (2013) Guest editorial: Theories and methods for advanced wireless relays; issue ii. *IEEE J Sel Areas Commun* 31(8):1361–1367. <https://doi.org/10.1109/JSAC.2013.130801>
- Kang JM, Kim IM, Kim HM (2017) Optimal training design for MIMO-OFDM two-way relay networks. *IEEE Trans Commun* 65(9):3675–3690
- Kong T, Hua Y (2011) Optimal design of source and relay pilots for MIMO relay channel estimation. *IEEE Trans Signal Process* 59(9):4438–4446
- Ma Y, Liu A, Hua Y (2014) A dual-phase power allocation scheme for multicarrier relay system with direct link. *IEEE Trans Signal Process* 62(1):5–16
- Nosratinia A, Hunter TE, Hedayat A (2004) Cooperative communication in wireless networks. *IEEE Commun Mag* 42(10):74–80
- Rong Y, Hua Y (2009) Optimality of diagonalization of multi-hop MIMO relays. *IEEE Trans Wirel Commun* 8(12):6068–6077
- Rong Y, Tang X, Hua Y (2009) A unified framework for optimizing linear nonregenerative multicarrier MIMO relay communication systems. *IEEE Trans Signal Process* 57(12):4837–4851
- Rong Y, Khandaker MR, Xiang Y (2012) Channel estimation of dual-hop MIMO relay system via parallel factor analysis. *IEEE Trans Wirel Commun* 11(6):2224–2233
- Sheng Z, Leung KK, Ding Z (2011) Cooperative wireless networks: from radio to network protocol designs. *IEEE Commun Mag* 49(5):64–69
- Soldani D, Dixit S (2008) Wireless relays for broadband access [radio communications series]. *IEEE Commun Mag* 46(3):58–66
- Wang R, Tao M (2012) Joint source and relay precoding designs for MIMO two-way relaying based on MSE criterion. *IEEE Trans Signal Process* 60(3):1352–1365
- Wang R, Tao M, Mehrpouyan H, Hua Y (2015) Channel estimation and optimal training design for correlated MIMO two-way relay systems in colored environment. *IEEE Trans Wirel Commun* 14(5):2684–2699
- Wang R, Mehrpouyan H, Tao M, Hua Y (2016) Channel estimation, carrier recovery, and data detection in the presence of phase noise in OFDM relay systems. *IEEE Trans Wirel Commun* 15(2):1186–1205
- Xie X, Peng M, Zhao B, Wang W, Hua Y (2014) Maximum a posteriori based channel estimation strategy for

two-way relaying channels. *IEEE Trans Wirel Commun* 13(1):450–463

Xu S, Hua Y (2011) Optimal design of spatial source-and-relay matrices for a non-regenerative two-way MIMO relay system. *IEEE Trans Wirel Commun* 10(5):1645–1655

Yu Y, Hua Y (2010) Power allocation for a MIMO relay system with multiple-antenna users. *IEEE Trans Signal Process* 58(5):2823–2835

Future Internet

► Smart Identifier Network

Future Internet Architecture

► Blockchain: Enabling Future Internet with Intrinsic Security

Future Networking

► Information-Centric Networking: Basic Principle and Architecture

Future Wireless Network Architecture and Network Slicing

Hang Zhang, Nimal Gamini Senarath, Xu Li, Ngoc Dung Dao, and Hamid Farmanbar
Huawei Technologies Canada Co., Ltd., Ottawa, Canada

Definitions

Network slicing is an artificial intelligence (AI) technique to automatically develop and deploy multiple virtual networks (network slices) utiliz-

ing shared physical resources of wireless infrastructure networks.

Introduction

Many factors are driving the redefinition and redesign of future wireless network architecture. These factors include new types of customer services, the huge disparate service QoS requirements, new format of general wireless network infrastructure (GWNI) with the integrated clouds, new requirements of rapid deployment of networks, and new demands on openness of network development and operation. The development of advanced techniques, such as network function virtualization (VNF) (Original NFV White Paper 2012; NFV White Paper #2 2013; NFV White Paper #3 2014), SDN (ONF White Paper 2012; OpenFlow White Paper 2008; Open Flow-enabled SDN 2014), AI, and big data, has made virtualization of network possible in the future wireless network. A virtualized network (VN) or a network slice (NGMN 2016; NGMN 5G White Paper 2014) is a collection of network resources including virtualized network functions (NF), to serve certain types of services. Network slicing techniques are techniques which create multiple virtual networks using a common physical GWNI. Network slicing techniques are expected to revolutionize the future wireless network architecture and have profound impact on wireless telecommunication industry.

Considering the aforementioned facts, the future wireless network architecture design should enable:

- Support of a wide variety of service types with huge disparate service requirements
- Full automation of slice creation, adaptation, and maintenances
- Guaranteed slice performance during the lifetime of a slice
- Simplified and scalable network control and management
- Rapid GWNI adaptation to performance degradation and traffic load changes

- Openness to allow customers to define their slices and even operate their slices
- Openness to allow third-party involvement in certain types of network control and management
- Future-proof architecture

Given the requirements of future wireless network architecture, the fundamental design philosophy of future network architecture should be the modularization of network control/management functions and the service-customized composition of these modules. This philosophy leads to the following three key design principles:

Principle 1: Separation of Resource Assignment (Logical or Physical)-Related Functions from Other Functions

Resource assignment/allocation includes two levels:

- Slice level: define resource allocation/assignment to slices, and manage the multiplexing of the infrastructure network resource among multiple slices.
- End points (devices/servers of a slice) level: manage multiplexing of slice resources among multiple end points.

Other network control/management functions are not directly involved in the allocation/assignment of resources but provide the information required by the resource control/management functions.

Such a separation allows slice providers to retain exclusive control over resource allocation while allowing third parties to provide other functions. A slice provider provides network slices to slice customers and is a new and key role in future wireless network industry.

Principle 2: Object-Oriented Functions Classification

Network control/management functions need to be identified and classified based on the types of objects which are controlled and managed. Their objects include:

- End points of a slice (devices/servers): end point's location and activity state need to be monitored, and this information is needed for traffic packet delivery.
- Infrastructure equipment elements: the elements, including radio nodes, clouds (a DC can be abstracted as a network element with certain process capability), and transport network equipment, need to be configured. Their performance needs to be monitored, and this information is used for resource control and management.
- Operation data: data logged during network operation is raw data from which information, such as service performance, traffic load, and resource utilization, can be extracted. The operation data can be used for various purposes, e.g., infrastructure network performance optimization, slice performance assurance, etc.
- Composed/deployed slices: a deployed slice can be viewed as a network product whose performance and behavior need to be continuously monitored, and the status statistics are needed for slice performance assurance.
- Media content: media content can be created by content producers and shared by remote parties using wireless networks. Caching and sharing of media content within networks need to be managed to ensure efficient content cache and delivery.

Such a classification enables a systematic design and definition of clear scope and responsibility for each function class.

Principle 3: Reorientation of Network Control/Management Function Classes as Services

Reorientation of control/management function classes as services enables a service-based architecture (SBA). Such an architecture makes it easy to define interactions among these services and, furthermore, support openness of future wireless network development, operation, and management.

In this paper, the function classes which provide information for resource control/management services are referred to as network operation assistance services (NOAS).

Network Control and Management Services

Based on the design principles, the following network control/management services are identified and classified. These function classes are briefly described in paper “Future wireless network: MyNET and SONAC” (Zhang et al. 2015).

Resource Control and Management Services

There are two resource control and management services: service-oriented virtual network automatic creation (SONAC)-Com service for slice composition and SONAC-Op service for slice operation.

- Slice composition service – SONAC-Com service

SONAC-Com is an intelligent network slicing technique that utilizes advanced techniques such as AI in wireless networks. The key responsibility of SONAC-Com is to rapidly develop and deploy end-to-end service-customized slices based on customer service requirements using a common infrastructure resource pool. SONAC-Com is designed to enable full automation of slice life cycle control and management. SONAC-Com should have a hierarchical architecture, and its functions need to be tailored for different technical domains, such as RAN domains, transport network domains, and core network domains. The functions at different level of the hierarchy have different responsibilities. The SONAC-Com function at top level determines the requirements of domains in lower level (e.g., RAN domains, transport domains, and core network domains) based on end-to-end service requirement. The top-level SONAC-Com also coordinates resource

assignments proposed by domain SONAC-Com functions. Regarding resource allocation and configuration, slice definition includes three key aspects:

- Software Defined Topology (SDT) determines the needed network functions (NF) and the placement of NFs in the infrastructure network, e.g., clouds in core and in RAN. SDT also determines the end-to-end logical topology of a slice, i.e., logical links among these NFs and capacity requirement of each logical link.
- Software Defined Resource Assignment (SDRA) maps the logical topology to the infrastructure network. For example, SDRA in transport domain needs to map logical links between NFs in core domain and in RAN domains to transport networks.
- Software Defined Protocol (SDP) determines data transmission protocols. The end-to-end protocols (e.g., TCP or non-TCP) is determined by SDP at top level. SDP functions are tailored to different technical domains. For example, SDP in RAN domain determines the service-customized access link protocols, i.e., selected protocol elements and their placement in RAN network.

After the above design procedure, a slice can be deployed by configuring the infrastructure network equipment elements and instantiating and/or configuring NFs in clouds utilizing the techniques developed, e.g., by ETSI-NFV.

SONAC-Com is also responsible for on-demand slice adaptation based on changes in slice traffic load, performance of slice, and/or infrastructure resource pool size.

- Slice operation service – SONAC-Op service
- After the deployment of a slice, the traffic packets of authorized devices/servers of the slice are delivered over the slice. It is the responsibility of SONAC-Op to associate devices/servers with slices and to operate slices in order to provide end-to-end connectivity services. The provisioning of an end-to-end connectivity service by SONAC-Op, using the allocated slice resource, can

include three aspects: (1) determination of the packet delivery logical path from source device(s)/server(s) to destination device(s)/server(s), (2) determination of the mapping between the logical path and physical network, and (3) determination of protocol of the end-to-end connection, including end-to-end protocol and per link protocol, if on-demand and dynamic protocol determination is required. Similar to SONAC-Com, SONAC-Op includes functions distributed in domains in a hierarchical architecture. The functions of SONAC-Op are tailored to different technical domains. SONAC-Op function in core domain determines the logical path of packet delivery in core network. SONAC-Op functions in access link of RAN network nodes are responsible for real-time resource assignment to devices based on policies defined by SONAC-Com.

NOAS Services

NOAS functions are identified and classified below.

- Data log and Analytics Management – DAM service

DAM service controls and manages operation data. The key responsibility of DAM is on-demand data log and analytics. DAM is the only service that has the capability to log raw data and provide a unified log and analytics service to other NOAS services and authorized third parties.

- Infrastructure network Management – InfM service

InfM service controls and manages infrastructure network elements. The future infrastructure network resources should be “elastic” to adapt to the changes in traffic loads and performance of infrastructure networks. In order to enable network slicing, the available resources of infrastructure networks need to be abstracted and provided to SONAC-Com on an on-demand basis. The InfM functions are tailored to different technical domains. For example, InfM functions in RAN network are

responsible for on-demand powering on/off of radio network nodes. InfM needs service from DAM and provides its service to SOANC-Com and authorized third parties.

- Connectivity Management – CM service

CM service controls and manages reachability of devices. The key responsibility of CM service is to provide reachability information of devices for the purpose of traffic delivery over slices. The reachability includes two sets of information – where and when a device is available. The device location tracking/resolution function tracks device location and provides the information regarding where to deliver packets. The device activity configuration and tracking function configures and tracks the activity status of a device and provides information regarding when to deliver packets to a device. CM should be designed to fully utilize certain service-specific characteristics, such as the predictable patterns in downlink traffic packet transmission or location of devices. Thus CM functions can be customized to different services. CM should be designed to enable the decoupling of device name and its location, which allows device name to remain unchanged despite changing in its location. CM service has a hierarchical architecture that is essential to enable efficient routing. CM provides service to SONAC-Op and authorized third parties.

- Customer Service Management – CSM service

CSM service manages deployed slices. A customer slice can be viewed as a “product” of networks. After such a slice is deployed, (1) the rule of authorization of slice access needs to be maintained for controlling the use of the slice, (2) the performance of the slice needs to be assured during the entire life cycle of the slice, (3) a tight matching between traffic load and resources of the slice should be maintained, (4) usage of the slice by authorized devices/subscribers should be monitored for the purpose of charging, and (5) characteristics of traffic data of the slice and devices behaviors should be monitored for better operation of the slice. CSM service is

designed for aforementioned purposes. CSM service utilizes the DAM service to obtain the required information and provides its service to SONAC-Com for customer slice adaptation, to SONAC-Op for slice operation optimization, and to slice providers and authorized parties for charging purposes.

- Cache and Forwarding Management – CFM service

CFM service controls and manages media content. In the future, sharing media content using wireless networks will be one of the most popular types of services. A slice provider can define a dedicated media content cache and forwarding slice to enable the efficient sharing of media content. CFM service manages the media content registration, content cache, and cache access and forwarding. CFM service needs content-access-related information provided by DAM services for its operation. CFM provides information such as “destinations” of media content for caching purposes and “sources” of media content for content access purposes.

These NOAS services are deployed as slices since each NOAS service consists of functions distributed over infrastructure networks and interconnected.

Future Wireless Network Architecture

Given these identified network control/management services, one example of potential future wireless network architecture is shown in Fig. 1.

As shown in Fig. 1, SONAC-Com service is implemented as a slice. SONAC-Com is the most intelligent service in the overall architecture and is programmed by slice providers. SONAC-Com is capable of automatically and rapidly developing and deploying other slices. The resource pool that SONAC-Com uses to deploy slices is the generic wireless network infrastructure resource pool which includes all hard (e.g., clouds and telecommunication network equipment) and soft (e.g., spectrum) infrastructure resources. Based

on the requirements and descriptions of services, SONAC-Com defines slices and provides instruction on deployment of slices.

In this architecture, all the network control/management services and customer services are viewed and treated in the same way by SONAC-Com. This makes scalable and systematic network control and management possible. Such architecture naturally enables openness of network operation and management.

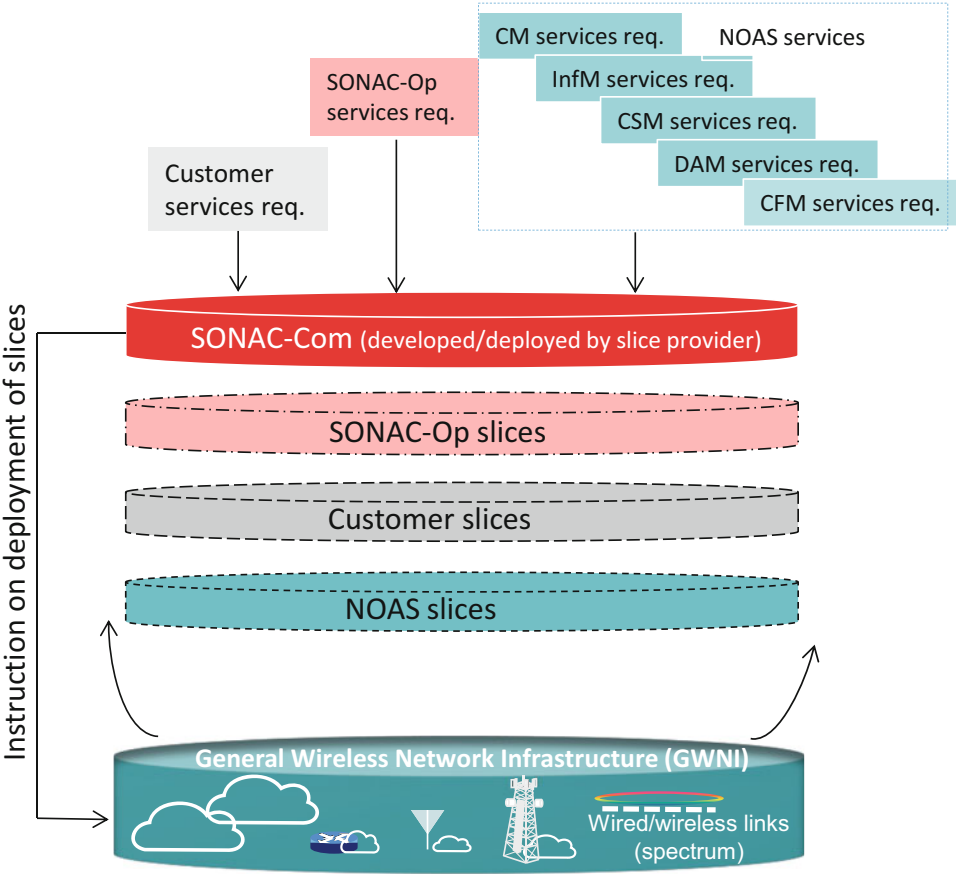
Development and Deployment of Slices

Development and Deployment of SONAC-Com Service Slice

SONAC-Com functions are developed and deployed by slice providers. After this procedure, other slices can be automatically developed by SONAC-Com as shown in Fig. 1. In order for SONAC-Com to develop other slices, SONAC-Com needs to maintain libraries of network functions (NF) designed for different technical domains and services. Each domain SONAC-Com function maintains the information describing its corresponding infrastructure networks. The information contains the specifications of physical network nodes, clouds, physical links, and topology. Given requests from slice customers, available infrastructure resources, and NF libraries, SONAC-Com service would be able to create all other slices.

Development and Deployment of SONAC-Op Service Slice

SONAC-Op slice is developed and deployed by the execution of the programmed SONAC-Com functions. This process includes (1) obtaining the SONAC-Op service requirement from slice providers; (2) deciding SONAC-Op function selection for different technical domains, the placement of SONAC-Op functions, the connections among these functions, and the policies that selected SONAC-Op functions need to follow during operation; and (3) deploying the SONAC-Op slice. After the deployment of SONAC-Op slice, the connections between SONAC-Com and



Future Wireless Network Architecture and Network Slicing, Fig. 1 Example of potential future wireless network architecture

SONAC-Op need to be established. A shared SONAC-Op slice can operate multiple customer slices. Customer-dedicated SONAC-Op slice can also be developed and deployed, if requested, at the same time a customer user plane (UP) slice is deployed.

Development and Deployment of NOAS Service Slices

The process of development and deployment of NOAS service slices is the same as that of a SONAC-Op service slice. After the deployment of NOAS slices, the required connections need to be established between NOAS slices and the SONAC-Op slice. The required connections also need to be established between NOAS slices. NOAS slices can be deployed before customer requests arrive. At the time customer slices are

deployed, these NOAS slices can be activated and/or configured. Customer-dedicated NOAS slices can also be developed and deployed, if requested, at the same time a customer user plane (UP) slice is deployed.

Development and Deployment of Customer Slices

A customer UP slice includes user plane (UP) NFs. The process of development and deployment of customer UP slices is the same as that of a SONAC-Op slice. Customers can define their private NF library and can even indicate the preferred placement of NFs. Some customer services may need UP plane slice, dedicated NOAS slices, and/or dedicated SONAC-Op slice. After a UP slice is deployed, connections need to be

established between the UP slice and SONAC-Op slice.

A customer slice can be deployed upon request. However, some purpose-specific UP slice can be deployed by slice providers prior to any customer request. A slice can be either dedicated to a single customer or shared among multiple customers if their service attributes and requirements are similar. For example, all of smart reading customers can share one UP slice – smart reading slice – and data traffic of these customers can be treated in the same way. Considering this scenario, when a service request is received, SONAC-Com needs (1) to determine which purpose-specific UP slice would be suitable for this customer, (2) to perform service admission control to make sure that the estimated resource requirement can be satisfied without causing any negative impact on other existing customers, and (3) to configure the corresponding SONAC-Op and NOAS functions.

Slice Resource Assignment Formats

There could be multiple formats of resource allocation to a slice.

- Format A (well-defined slice): an end-to-end slice with (1) definition of NFs and placement of the NFs in the infrastructure networks; (2) slice logical topology, i.e., logical links among NFs, including the links between NFs and RAN nodes; (3) capacities of logical links of this slice; (4) mapping of logical links to infrastructure networks; and (5) definition of end-to-end protocols and per logical link protocols.
- Format B (moderately defined slice): similar to Format A. However, there is no (3) definition of capacities of logical links of the slice.
- Format C (loosely defined slice): slice with only (1) definition of NFs and placement of NFs in the infrastructure network.

Variations of these formats are also possible. The format should be selected based on service attributes, e.g., the degree of statistical predictability of traffic load pattern, the degree of device mobility, and the service quality

requirement. The network load and congestion control is another factor which impacts on the selection of format. Since Format A defines logical link capacities, the network-level traffic congestion control can be managed at the slice level by SONAC-Com, which “sees a bigger picture” of load distribution over infrastructure networks. Furthermore, complexity balance between SONAC-Com and SONAC-Op may also need to be considered. To define a well-defined slice requires a higher level of intelligence from SONAC-Com. However, a Format A slice enables the relative simple process of SONAC-Op, which results in shorter process latency and less signaling overhead. Format C requires more intelligent SONAC-Op process but less complicated SONAC-Com process.

Operation of a Deployed Slice

After a slice is deployed, the slice is ready to deliver packet traffic. Different functions of SONAC-Op can be designed for different slice formats.

Format A: After a well-defined end-to-end customer UP slice is deployed, SONAC-Com needs to provide slice information to SONAC-Op. A Format A slice enables a slice-level forwarding rule, i.e., slice-level routing table, to be pre-configured. Therefore the procedure of per end point (device/server) end-to-end path establishment can be simplified. Per device uplink traffic packet delivery becomes the selection of the right “entry point” of such a slice by RAN access nodes. The downlink data packet delivery to a mobile device becomes the selection of the right “exit point” of such a slice by network. The NFs in UP can perform a virtual router function. The responsibility of SONAC-Op for the purpose of packet delivery is to control and to manage the end point forwarding tables at virtual routers utilizing CM services. Such a packet delivery approach can be viewed as a packet-oriented scheme over virtual networks. Format A slice definition together with virtual router-based packet delivery approach combines the circuit-oriented technique at slice level and the packet-oriented technique at packet level. Furthermore,

a Format A slice designed for customers, such as vertical industry customers who collect data from their devices or control the behaviors of all these devices, can significantly simplify the operation of slice. In this case, SONAC-Op may not need to perform path management at per device level since such a type of service does not need to differentiate between devices. For operation of Format A slices, SONAC-Op needs to control resources allocation among slices based on the capacity requirements of logical links of slices in order to ensure slice-level QoS.

Format B: The responsibility of SONAC-Op in Format B slices is similar to the responsibility in Format A slices. However, since there is no definition of capacities of logical links, slice-level performance assurance may be more challenging. For example, if a radio node supports two slices, SONAC-Op at access link of the radio node – acting as access link scheduler – may under-provision or over-provision resources to these slices.

Format C: The responsibility of SONAC-Op in Format C slices may need to perform on-demand configuration of interconnection of NFs based on the dynamic service requirement, in addition to those of Format B. SONAC-Op also needs to control the per device end-to-end path establishment.

The interaction between SONAC-Op and other slices are illustrated in Fig. 2.

As shown in Fig. 2, CM NOAS service provides device reachability information to SONAC-Op for packet delivery path control and management. CSM NOAS service provides behavior context information of devices of a slice

to SONAC-Op for slice operation optimization. SONAC-Op performs slice operation based on the real-time information provided by NOAS services and pre-configured policies.

Adaptation of Deployed Slices

SONAC-Com is programmed to automatically and rapidly modify existing slices. There is a number of factors that may trigger slice adaptation.

- Customer service assurance aspect

Slice performance is monitored by functions of CSM service with assistance of DAM service after a slice is deployed. If CSM identifies abnormal events, such as a mismatching between resources and traffic load of a slice, CSM service will trigger a slice adaptation.

- Infrastructure resource optimization aspect

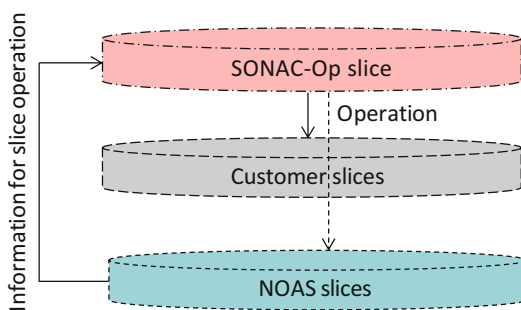
Performance assurance of infrastructure networks and optimization of infrastructure resource utilization can result in changes in infrastructure network topology and capacity. This can cause the adaptation of existing slices. InfM can identify these changes with assistance of DAM service and trigger a slice adaptation.

- Device service aspect

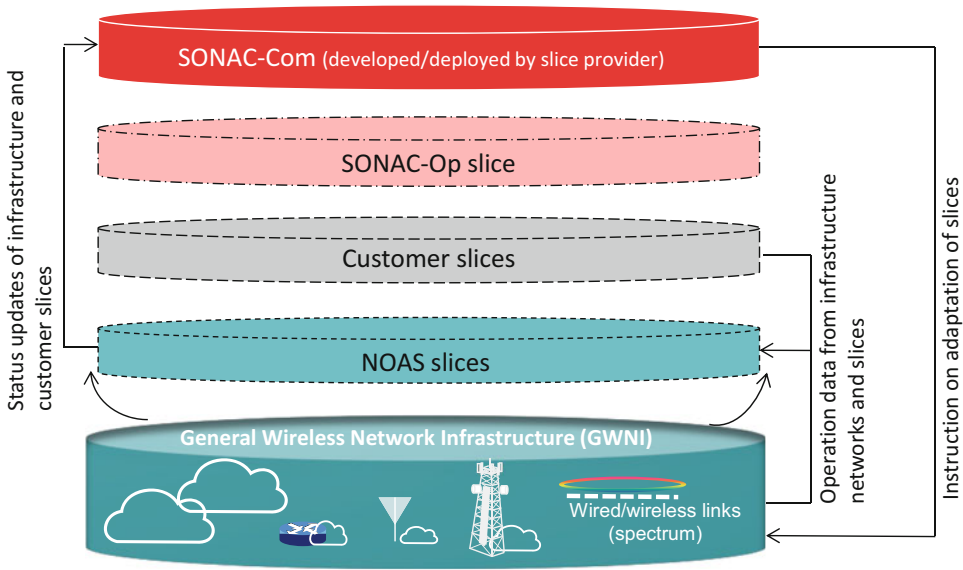
For some extreme cases where a customer slice is dedicated to one customer with a group of co-located mobile devices, e.g., an emergency ambulance customer, some service-specific NFs need to be deployed close to the edge of the network and need to migrate along with the mobile devices. In this case, the slice topology may be adapted to the movement of the mobile devices.

- Customer service aspect

For some services that utilize deployed sensors to monitor changes in, e.g., temperature, in order to detect occurrence of certain incidences, a dedicated slice can be deployed for such a service. The occurrence of certain incidences may lead to spikes in traffic. This requires SONAC-Com to rapidly modify the slice resources to adapt to the change in traffic load. Such incidences can be detected



Future Wireless Network Architecture and Network Slicing, Fig. 2 Slice operation



Future Wireless Network Architecture and Network Slicing, Fig. 3 Adaptation of slices

by, e.g., service application functions. In this case, application functions may trigger a slice adaptation.

Slice adaptation and interaction among SONAC-Com and other NOAS slices are illustrated in Fig. 3.

As shown in Fig. 3, NOAS services, such as CSM and InfM services, obtain information of performance of slices and infrastructure networks by utilizing the service of DAM. CSM and InfM services provide status reports to SONAC-Com when certain pre-configured conditions are met. SONAC-Com makes decision on slice adaptation based on the information and policies pre-configured by slice providers.

Openness and New Business Model Support

In future wireless network industry, a new business model can be expected where multiple types of players will be involved in supporting communication services. These players could be:

- Slice providers who provide end-to-end slices by using the integrated generic infrastructure resource provided by infrastructure providers
- Slice customers (owners) who request and obtain end-to-end slices from slice providers and may control/manage the slices to provide further services to their customers (consumers), e.g., subscribers
- Slice customers (consumers) who consume slice resource to obtain communication service, e.g., mobile subscribers

The design philosophy discussed in this paper enables flexible distribution of network control/management services/functions among players, i.e., these control/management services in the discussed architecture can be provided by slice providers or by other players such as authorized third parties. This architecture also makes it possible to share various status information with authorized third parties. For example, DAM service can provide on-demand information exposure services to authorized third parties. Utilizing DAM service, CSM service can provide slice status information to customers of slices, and InfM service can provide information of infrastructure network resource and traffic load to authorized third parties.

This new business model needs new business revenue support. In the discussed architecture, NOAS services, such as CSM and DAM services, can provide sufficiently detailed and granulated information such that charging information can be derived.

Conclusion

Network virtualization and automation are expected to have profound impact on the wireless network industry. In the future, advanced techniques, such as AI and big data, will become essential techniques for network virtualization and slicing. With the development of these technologies and with a better understanding of future service demands, the future wireless networks will be designed to provide true ubiquitous connectivity services to all types of customers and will benefit all walks of life and whole society.

Cross-References

- [Big Data in 5G](#)
- [Wireless Edge Caching](#)

References

- NFV White Paper #2, October 15–17, 2013. http://portal.etsi.org/NFV/NFV_White_Paper2.pdf
- NFV White Paper #3 October 14–17, 2014. http://portal.etsi.org/NFV/NFV_White_Paper3.pdf
- NGMN 5G Project Requirements & Architecture–Work Stream E2E Architecture, Description of Network Slicing Concept Version 1.0, 13th January 2016, by NGMN Alliance
- NGMN 5G White Paper – Executive Version, Version 1.0, December 22, 2014. http://www.ngmn.org/uploads/media/141222_NGMNExecutive_Version_of_the_5G_White_Paper_v1_0_01.pdf
- ONF White Paper, Software-Defined Networking: The New Norm for Networks, April 13, 2012. <https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf>
- Open Flow-enabled SDN and Network Functions Virtualization, February 17, 2014. <https://www.opennetworking.org/images/stories/downloads/sdn-resources/solution-briefs/sb-sdn-nvf-solution.pdf>
- OpenFlow White Paper: Enabling Innovation in Campus Networks, March 14, 2008. <http://archive.openflow.org/documents/openflow-wp-latest.pdf>
- Original NFV White Paper, October 2012. http://portal.etsi.org/NFV/NFV_White_Paper.pdf
- Zhang H, Vrzić S, Senarath G, Đào N-D, Farmanbar H, Rao J, Peng C, Zhuang H (2015) Future wireless network: MyNET and SONAC. *Netw Mag IEEE* 29(4):14–23

Game Model-Aided Cooperative Spectrum Access

► [Game Theory-Assisted Cooperative Spectrum Access](#)

Game Theoretic Modeling of Zero-Rating Internet Provisioning

Masoud Asghari¹, Saleh Yousefi¹, and Dusit Niyato²

¹Department of Computer Engineering, Urmia University, Urmia, Iran

²School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore, Singapore

Synonyms

[Free basics](#); [Internet.org](#); [Sponsored content](#); [Sponsored data](#); [Zero-rating](#)

Definitions

Zero-rating Internet is the concept of allowing end-users to access the basic services of some websites via mobile network operators (MNOs) without charge. More specifically, the MNO does

not subtract data usage of zero-rated services from the end-users' data quota.

Sponsored content (i.e., sponsored data) is a similar context to zero-rating, where the end-users still get access to some websites and contents for free or with a discount. However, unlike zero-rating, in sponsored content the websites pay the cost of the sponsored contents' bandwidth to the MNOs.

A *zero-rating platform* is a relaying service which allows websites to easily make their content available as zero-rating through the platform.

Historical Background

Websites such as Facebook, Google, and Wikipedia have been providing zero-rating access of their services (e.g., Facebook Zero, Google Free Zone, Wikipedia Zero) with partnership of some mobile network operators (MNOs) in several less developed countries. Their primary aim is to provide free basic access of their services for people in less developed countries and show them the true value of Internet services (de Miera Berglind 2016). Most of the works related to zero-rating are focused on regulations as in de Miera Berglind (2016) and Sambuli (2016) and the assessment of its benefits and costs as in Futter and Gillwald (2015) and Eisenach (2015).

In addition to zero-rating services of particular websites, zero-rating platforms are provided

by some content and service providers. An example of a zero-rating platform, launched in August 2013, is *Free Basics* (formerly called Internet.org) that is provided by Facebook to bring affordable access of selected Internet services to the people of less developed countries. When an MNO becomes a partner with this platform, the MNO allows its customers to receive basic free access to websites that are included in the platform. As a response to the concerns over net neutrality, in May 2015, Facebook announced the platform as an open program for developers to create services that integrate with the platform. Consequently, *Free Basics* allows websites to provide their basic services through the platform for free on condition that they meet the basic technical requirements of the platform to make a simplified version of their website that is data-efficient (e.g., services that use VoIP, video, file transfer, or large photos are not compatible). Neither the platform nor the participating websites pay the MNOs for the data that the end-users consume through *Free Basics* (Int 2017).

Technically speaking, *Free Basics* works similar to a virtual private network (VPN) or a proxy service. Partner MNOs do not charge their end-users' traffic to/from *Free Basics* servers. End-users of the partner MNOs can connect to *Free Basics* servers, and afterward their requested websites are relayed by the service. Once connected to *Free Basics*, the end-user is only allowed to see a simplified version of the participating websites at the platform. An application and a landing page are provided by *Free Basics* that show a list of available websites to the end-users.

Currently, different bundle pricing schemes are commonly employed by many MNOs. In these schemes, usually a limited amount of voice calls and data during an interval are sold to the end-users as a prepaid bundle. Moreover, usage-based pricing is imposed for unbundled and overage calls and data (Sen et al. 2013). A bundle pricing scheme for the MNO with inclusion of zero-rating services in the bundles is proposed in Asghari et al. (2017), and the circumstance is

modeled as a Stackelberg game with an MNO and its end-users as the players of the game where the MNO decides about provisioning of zero-rating service in its bundles as well as the bundles prices and the end-users respond by deciding about buying the bundles. The proposed model assists the MNO to make a decision on including zero-rating services in its bundles as well as appropriate pricing scheme. Numerical studies show that depending on the end-users' preferences, the MNO may increase its profit by including zero-rating services in its bundles (Asghari et al. 2017).

Recently, the context of sponsored content has attracted lots of attention for its economic challenges. Unlike the zero-rating concept that the only profit source of the MNO is from the end-users, in the sponsored content concept, the MNO gets payments from the content providers in addition to the end-users, while the content providers pay a fraction or complete cost of the sponsored contents' bandwidth to the MNO instead of the viewers to attract more viewers. The work of Andrews et al. in (2013) can be considered as a fundamental work on economic aspects of sponsored content. They considered a Stackelberg game with an MNO and a content provider, where the MNO sets a pricing schedule to the content provider for sponsoring and the content provider responds by deciding how much content to sponsor. They concluded that a coordinating contract can be designed aiming at maximizing total system profit. A methodology in Andrews et al. (2014) is developed for extracting required parameters of the model in Andrews et al. (2013). Later in Jin et al. (2015), Jin et al. extended the model of Andrews et al. (2013) even further by including multiple content providers and considering different versions of the model. Additionally, Zhang and Wang studied the impact of sponsored data strategy on content providers with different scales (Zhang and Wang 2014). Finally, Joe-Wong et al. considered content providers' choices and introduced a framework for content providers to decide which and how much content to sponsor (Joe-Wong et al. 2015).

Foundations

A zero-rating platform such as *Free Basics* is unable to provide any services to the end-users unless the end-users can access it through the MNOs for free. However, the MNOs only grant a free access to zero-rating platforms if the MNOs believe that it will be beneficial for them. Considering that in current setup of zero-rating platforms no monetary reward is transferred from the participating websites and zero-rating platform providers to the partner MNOs, the MNOs seek for the extra profit in this partnership from end-users side. Thus, the MNOs face a question of whether to cooperate with zero-rating platforms or not.

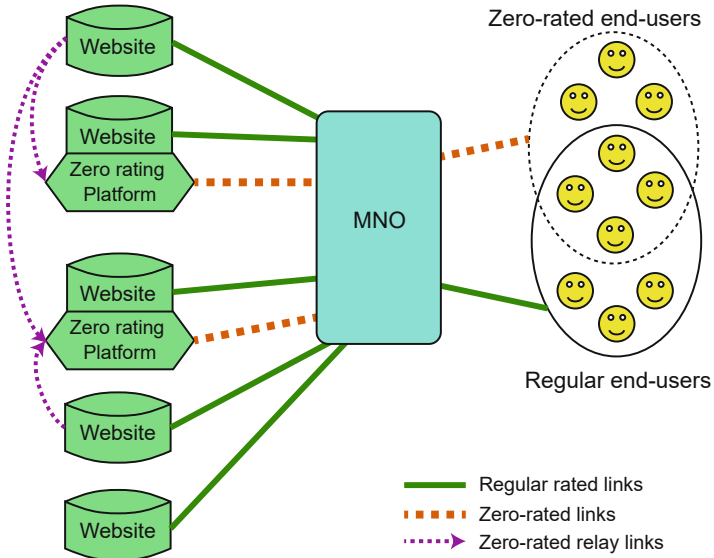
The interaction between an MNO and its end-users toward providing their access to zero-rating platforms' services can be modeled by a variant of Stackelberg game where a bundle pricing scheme of the MNO is considered. Thereby, the effects of including zero-rating platforms' services in the bundles can be studied (Asghari et al. 2017). The results of the game assists the MNO to react to this new trend and make a decision on provisioning of zero-rating services in its bundles as well as determination of appropriate pricing strategies.

Figure 1 shows the system model and its major players. A MNO and its end-users are considered as the active players of a game where each end-user may prefer regular services, zero-rating services, or both. All websites are available for the end-users through regular rated links provided by the MNO. Meanwhile, some partially opened or closed zero-rating platforms are presented, and they essentially use zero-rated links brought by the MNO to communicate with its end-users. In this context, websites can use zero-rating platforms as relays to present their services as zero-rated to the end-users.

Considering an MNO in a region, it faces with a question of whether it should cooperate with zero-rating platforms and offer zero-rating services to its customers or not. In case of offering zero-rating services, the MNO may change the prices to compensate the cost of zero-rated traffic and substitution effects of zero-rating services on non-free data usage.

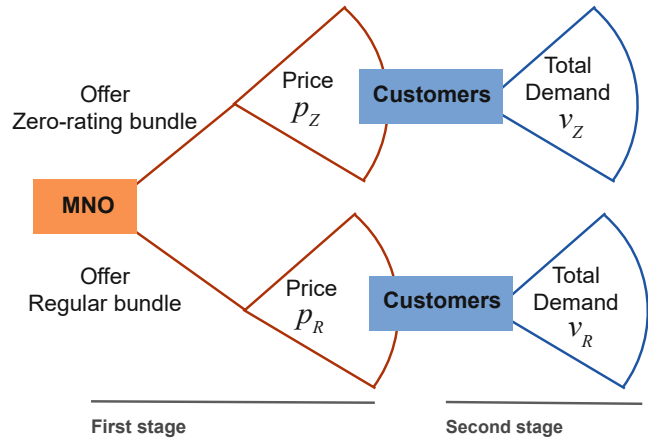
To model the circumstance as a game, it is assumed that the MNO offers a bundle with limited amount of voice calls and data during an interval. The bundle may include zero-rating platforms' services or not, and the MNO only provides the bundle either with or without zero-rating services for all of the customers. The

Game Theoretic Modeling of Zero-Rating Internet Provisioning, Fig. 1 Zero-rating system model



Game Theoretic Modeling of Zero-Rating Internet Provisioning,

Fig. 2 Extensive form representation of the game (players: MNO and customers)



bundle without zero-rating services is called “regular bundle,” and the bundle with zero-rating services is called “zero-rating bundle” where the zero-rating bundle includes every services of the regular bundle, but its price may be different. A simple example of a regular bundle can be considered as prepaid minutes of call package including data quota or without data.

The situation can be modeled as a variant of Stackelberg game which is a sequential game, where the MNO is the leader and moves first and the customers are the followers. It is assumed that the zero-rating platforms are formed and they are willing to cooperate with the MNO. Figure 2 shows the extensive form representation of the game which consists of two stages. At the first stage, the MNO chooses to offer a bundle with zero-rating platforms’ services included or with the regular services alone, along the prices of the bundles (p_Z and p_R). At the second stage, the customers observe the MNO’s decision and choose whether they want to buy the offered bundle or not which forms the total demand of the customers (v_Z and v_R). The game is solved with backward induction where the followers decide how many bundles they buy considering the decisions of the leader. The leader can anticipate the follower’s best response and choose the prices and include zero-rating platforms’ services to maximize its profit. The solution of the game is a subgame perfect Nash equilibrium (SPNE) and contains the best strategies of the players, where

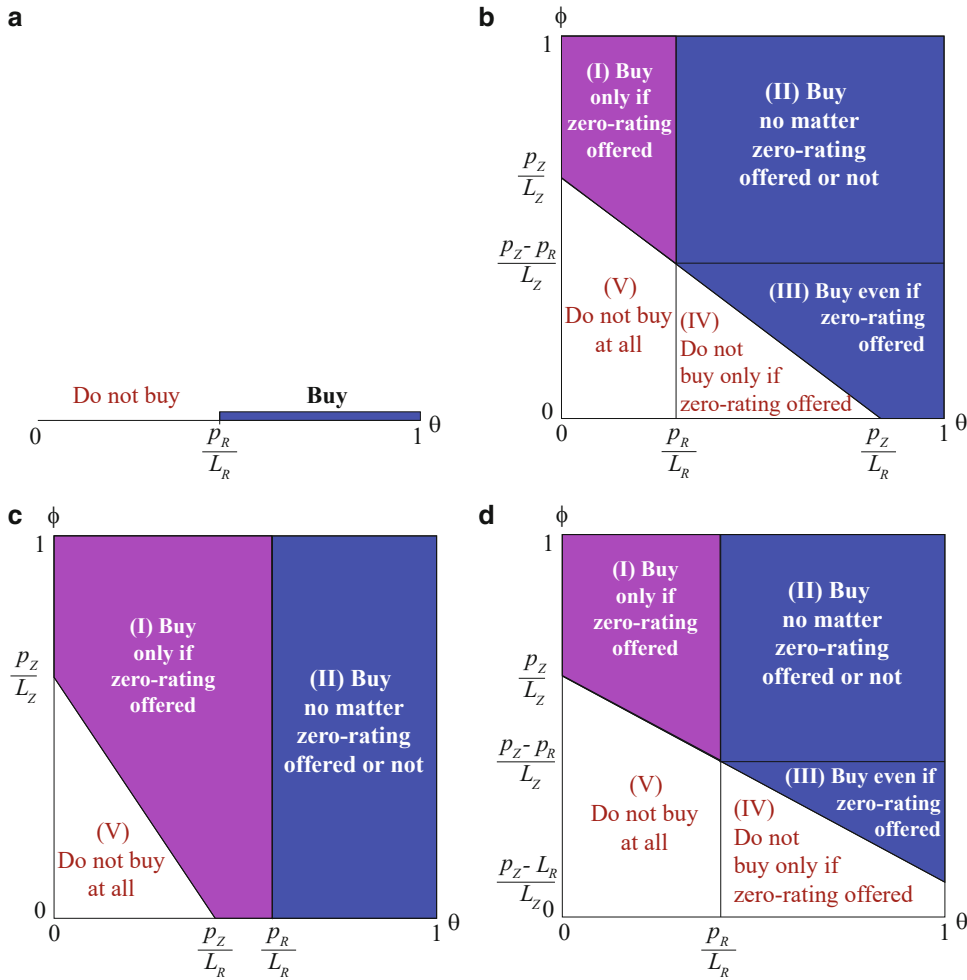
no player can benefit by only changing its own strategy.

It is considered that there is v_0 potential customer in the region. Net utility function of customer i for buying a bundle is considered as follows:

$$U_i = \begin{cases} \phi_i L_Z + \theta_i L_R - p_Z & \text{Buys a zero-rating bundle} \\ \theta_i L_R - p_R & \text{Buys a regular bundle} \\ 0 & \text{Buys no bundle} \end{cases} \quad (1)$$

where ϕ_i and θ_i are customers’ preference parameters regarding the zero-rating services and a regular bundle, respectively. Both ϕ_i and θ_i are considered independent and uniformly distributed in $[0,1]$. In addition, $L_Z \geq 0$ and $L_R \geq 0$ are the highest satisfaction levels (i.e., valuation) that a customer may get by having zero-rating services and a regular bundle, respectively. Moreover, p_R is the price for a regular bundle, and p_Z is the price for a zero-rating bundle which includes the regular bundle services as well. Every customer i selects one of the three options in its utility function to maximize its utility based on its ϕ_i and θ_i , the offered prices p_R and p_Z , and the type of offered bundle (zero-rating or regular).

Figure 3 shows two preference parameters ϕ and θ and the way that customers decide on any of their three choices given their utility function defined in (1). Figure 3a shows the preference



Game Theoretic Modeling of Zero-Rating Internet Provisioning, Fig. 3 Customers sensitivity map for (a) regular bundle, (b) zero-rating bundle when $((p_Z < L_R) \& (p_Z < p_R))$, (c) zero-rating bundle when $((p_Z < L_R) \& (p_Z < p_R))$, and (d) zero-rating bundle when $(p_Z \geq L_R)$. (Source: Asghari et al. 2017)

map of the customers when the regular bundle is the only offered bundle. It is obvious that $\frac{p_R}{L_R} \leq 1$. Figure 3b–d show the preference map of the customers when the zero-rating bundle is the only offered bundle, while Fig. 3b, c show the map when $(p_Z < L_R)$, and Fig. 3d shows the map when $(p_Z \geq L_R)$. Figure 3c shows a special case of Fig. 3b where $(p_Z < p_R)$. For simplicity, it is only considered the case of $p_Z \leq L_Z$.

The maps in Fig. 3b–d are divided into different regions that are marked with (I) to (V). The division lines show indifference points regarding the consumer preferences on the line. The total

demand functions of the customers for different type of bundles are derived as follows.

When the MNO only offers the regular bundle, Fig. 3a is applicable, and the customers only choose between buying a regular bundle or not buying at all. A customer who has $\theta = \frac{p_R}{L_R}$ is indifferent between buying or not. Thus, v_0 multiplied by the segment from $\theta = \frac{p_R}{L_R}$ to 1 determines the total demand function of the customers who buy the regular bundle which is equal to:

$$v_R = v_0 \left(1 - \frac{p_R}{L_R} \right). \quad (2)$$

When the MNO only offers the zero-rating bundle, all customers in regions (I), (II), and (III) will buy it. The line that separates these regions with regions (IV) and (V) corresponds to the condition of $\phi = \frac{p_Z}{L_Z} - \theta \frac{L_R}{L_Z}$, and the total area of all regions is equal to one. Thus, v_0 multiplied by the combined area of regions (I), (II), and (III) determines the total demand function of the customers who buy the zero-rating bundle which is equal to:

$$v_Z = \begin{cases} v_0 \left(1 - \frac{p_Z^2}{2L_Z L_R}\right) & \text{Otherwise} \\ v_0 \left(1 - \frac{2p_Z - L_R}{2L_Z}\right) & \text{if}(p_Z \geq L_R). \end{cases} \quad (3)$$

Finally, there are customers that always buy the bundle either with or without zero-rating services (i.e., $Z \cap R$) which are placed at regions (II) and (III). Their demand can be shown as:

$$v_{(Z \cap R)} = \begin{cases} v_0 \left(1 - \frac{p_Z^2 - 2p_Z p_R + p_R^2 + 2p_R L_Z}{2L_Z L_R}\right) & \text{Otherwise} \\ v_0 \left(1 - \frac{p_R}{L_R}\right) & \text{if}((p_Z < L_R) \& (p_Z < p_R)) \\ v_0 \left(1 - \frac{2p_Z L_R - L_R^2 - 2p_Z p_R + p_R^2 + 2p_R L_Z}{2L_Z L_R}\right) & \text{if}(p_Z \geq L_R). \end{cases} \quad (4)$$

These demands are the best responses of the customers to the leader's decision. The MNO is aware of these demands, and it decides on the prices and the choice of offering zero-rating services or not in the bundle to maximize its profit. Considering these demands, profit of the MNO when it includes zero-rating services (shown with π_Z) and when it does not include zero-rating services (shown with π_R) can be written as:

$$\pi_Z = (p_Z - c_R - c_Z)v_Z - c_S v_{(Z \cap R)}, \quad (5)$$

$$\pi_R = (p_R - c_R)v_R, \quad (6)$$

where c_R is the average cost of the regular services at the bundle for the MNO and c_Z is the average cost of including zero-rating services at the bundle for the MNO. Moreover, c_S is the substitution cost of zero-rating services at the bundle to the profit of the MNO. In other words, c_S is imposed by those customers who always buy the bundle either with or without zero-rating services ($Z \cap R$). When the regular bundle is the only offered bundle, the customers in ($Z \cap R$) would buy the regular bundle. Further, when zero-rating bundle is the only offered bundle, they will still buy the bundle. However, in this case, they may consume less non-free data than regular bundle

due to availability and substitution effect of zero-rating services in the zero-rating bundle.

To suggest the MNO to offer zero-rating platforms' services with the bundle or not and choose the best price for the bundle, the games is solved as follows:

First, the optimal p_R that maximizes π_R is found. Because of $\frac{\partial^2 \pi_R}{\partial p_R^2} < 0$, the solution of $\frac{\partial \pi_R}{\partial p_R} = 0$ gives the optimal p_R as:

$$p_R^* = \frac{L_R + c_R}{2}. \quad (7)$$

Furthermore, the optimal π_R becomes:

$$\pi_R^* = v_0 \frac{(L_R - c_R)^2}{4L_R}. \quad (8)$$

To find optimal p_Z and π_Z , the solution of $\frac{\partial \pi_Z}{\partial p_Z} = 0$ is found as:

$$p_Z = \begin{cases} \frac{A_1 \pm \sqrt{A_2}}{3} & \text{if}((p_Z < L_R) \& (p_Z \geq p_R)) \\ \frac{A_3 \pm \sqrt{A_4}}{3} & \text{if}((p_Z < L_R) \& (p_Z < p_R)) \\ A_5 & \text{if}(p_Z \geq L_R) \end{cases} \quad (9)$$

where

$$\begin{aligned} A_1 &= c_R + c_Z + c_S, \\ A_2 &= (A_1)^2 + 6L_Z L_R - 3c_S L_R - 3c_R c_S, \\ A_3 &= c_R + c_Z, \\ A_4 &= (A_3)^2 + 6L_Z L_R, \\ A_5 &= \frac{L_R + 2L_Z + c_S + 2c_R + 2c_Z}{4} - \frac{c_S c_R}{4L_R}. \end{aligned}$$

Only those situations that $A_2 \geq 0$ are considered. For $(p_Z \geq L_R)$, we have $\frac{\partial^2 \pi_Z}{\partial p_Z^2} < 0$; thus the solution of $\frac{\partial \pi_Z}{\partial p_Z} = 0$ gives the optimal p_Z in (9). However, for $((p_Z < L_R) \& (p_Z \geq p_R))$, we have $\frac{\partial^2 \pi_Z}{\partial p_Z^2} < 0$ only at $p_Z > \frac{A_1}{3}$, which is the part that it is maximized. Likewise, $p_Z > \frac{A_3}{3}$ should be selected for $((p_Z < L_R) \& (p_Z < p_R))$ case. Therefore, we take both plus signs of $\pm$ in (9). In addition, p_Z can be eliminated from the conditions in (9) which transforms it into:

$$p_Z^* = \operatorname{argmax}_{p_Z \in \Psi} (\pi_Z), \quad (10)$$

where

$$\begin{aligned} \Psi = \left\{ \min \left(\max \left(\frac{A_1 + \sqrt{A_2}}{3}, p_R^* \right), L_Z, L_R \right), \right. \\ \min \left(\frac{A_3 + \sqrt{A_4}}{3}, p_R^*, L_Z, L_R \right), \\ \left. \max (\min (A_5, L_Z), L_R) \right\}. \end{aligned}$$

Note that the expression in (10) is basically a selection algorithm which chooses p_Z^* among six discrete points. By substitution of p_Z^* from (10) into (5), the optimal π_Z^* can be obtained.

In order to determine the MNO's optimal decision, *profit margin* of the MNO is defined as $\Delta = \frac{\pi_Z^* - \pi_R^*}{\pi_R^*}$ which indicates the ratio of changes in MNO's profit when it includes zero-rating services in the bundle to the profit when it does not include the zero-rating services in the bundle. The MNO includes zero-rating services in the bundle only if $\Delta \geq 0$. Finally, the determined p_Z^* , p_R^* and the decision to include zero-rating

services when $(\Delta \geq 0)$ are in the unique SPNE of the game.

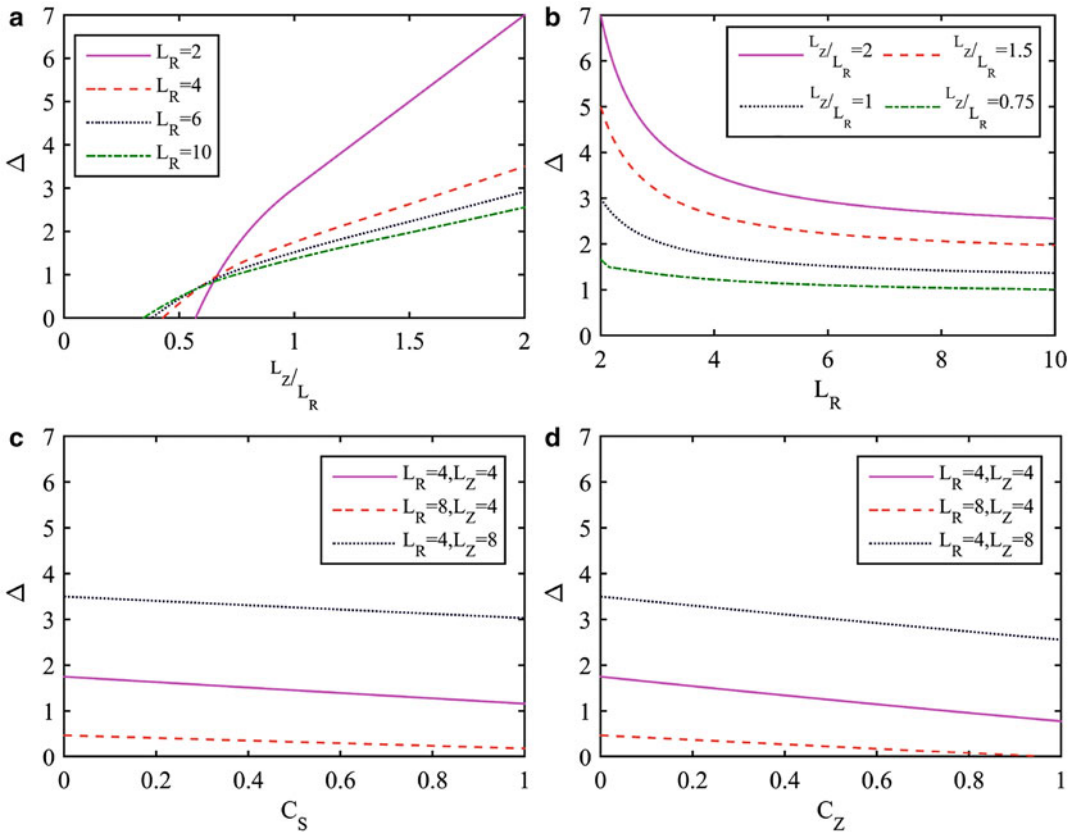
Figure 4 shows numerical examples. In the examples, we set $c_R = 1$, $c_Z = 0$, and $c_S = 0$ unless otherwise stated. Figure 4a shows $\Delta > 1$ (i.e., the MNO includes zero-rating services) with respect to the ratio of L_Z to L_R . Figure 4b is similar to Fig. 4a focusing on how L_R changes. It can be observed that as $\frac{L_Z}{L_R}$ increases, the profit margin of the MNO increases as well by offering zero-rating bundles. That is due to higher valuation of the end-users for zero-rating services. This effect weakens as L_R increases. Since, for small L_R , the MNO has low profit from regular bundles, and offering zero-rating bundles would have a greater impact on its profit. An important observation is that in most cases for $\frac{L_Z}{L_R} < 0.5$, the MNO is better off without zero-rating bundles. Therefore, the MNO tends to increase L_Z of the end-users by participating with zero-rating platforms that are more attractive for them.

Figure 4c, d show the impact of c_S and c_Z to the profit margin when all the other parameters remain the same. The increase in both of these costs reduces Δ which is evident, and c_Z is more effective than c_S because c_Z applies to all customers' demand. However, c_S only applies on demand of customers in $Z \cap R$ which are a portion of all customers. c_S can be interpreted as loss in out-of-bundle data sales caused by substitution effects of zero-rating services in the bundle.

In conclusion, the numerical examples show that depending on the costs and end-users preferences, the MNO can increase its profit by including zero-rating services in the bundle and choosing the bundle price in (10) for the cases of $\Delta > 1$.

Key Applications

As mentioned before, in zero-rating Internet solutions such as *Free Basics*, no monetary reward is transferred from participating websites and zero-rating platform providers to the MNOs. That is one of the major obstacles in provisioning of



Game Theoretic Modeling of Zero-Rating Internet Provisioning, Fig. 4 Numerical examples showing (a–b) positive Δ with optimal p_Z^* when $c_R = 1$, $c_Z = 0$, $c_S =$

0; (c) positive Δ with optimal p_Z^* when $c_R = 1$, $c_Z = 0$; and (d) positive Δ with optimal p_Z^* when $c_R = 1$, $c_S = 0$. (Source: Asghari et al. 2017)

zero-rating Internet that discourages the MNOs to participate in zero-rating practices. However, considering the provisioning of zero-rating Internet as a new opportunity for the MNOs to boost the sales of their services in the form of bundles by attracting more customers can make the provisioning of zero-rating services profitable for the MNOs and facilitate its provisioning. In addition, when multiple MNOs are considered, provisioning of zero-rating services can play an important role in their competition.

Cross-References

- [Dynamic Pricing](#)
- [Preferences and Utility Functions](#)

References

- Andrews M, Ozen U, Reiman MI, Wang Q (2013) Economic models of sponsored content in wireless networks with uncertain demand. In: 2013 IEEE conference computer on communications workshops (INFOCOM WKSHPs). IEEE, pp 345–350
- Andrews M, Bruns G, Lee H (2014) Calculating the benefits of sponsored data for an individual content provider. In: 2014 48th annual conference on information sciences and systems (CISS). IEEE, pp 1–6
- Asghari M, Yousefi S, Niyato D (2017) A mobile network operator's decision for partnership with zero-rating internet platforms. In: 2017 IEEE global communications conference (GLOBECOM). IEEE, pp 1–6
- de Miera Berglind OS (2016) The effect of zero-rating on mobile broadband demand: an empirical approach and potential implications. *Int J Commun* 10:18
- Eisenach JA (2015) The economics of zero rating. NERA Economic Consulting. <http://www.nera.com/content/dam/nera/publications/2015/EconomicsofZeroRating.pdf>

- Futter A, Gillwald A (2015) Zero-rated internet services: What is to be done?. *Broadband 4 Africa* 1:1–10
- Int (2017) Internet.org. <http://internet.org/>
- Jin Y, Reiman MI, Andrews M (2015) Pricing sponsored content in wireless networks with multiple content providers. In: 2015 IEEE conference on computer communications workshops (INFOCOM WKSHPs). IEEE, pp 1–6
- Joe-Wong C, Ha S, Chiang M (2015) Sponsoring mobile data: an economic analysis of the impact on users and content providers. In: 2015 IEEE conference on computer communications (INFOCOM). IEEE, pp 1499–1507
- Sambuli N (2016) Challenges and opportunities for advancing Internet access in developing countries while upholding net neutrality. *J Cyber Policy* 1(1):61–74
- Sen S, Joe-Wong C, Ha S, Chiang M (2013) A survey of smart data pricing: past proposals, current plans, and future trends. *ACM Comput Surv (CSUR)* 46(2):15
- Zhang L, Wang D (2014) Sponsoring content: motivation and pitfalls for content service providers. In: 2014 IEEE conference on computer communications workshops (INFOCOM WKSHPs). IEEE, pp 577–582

Game Theory

- [Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future](#)

Game Theory Based Privacy Preservation

- [Privacy Game](#)

Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future

Xiangyu Chen and Ka-Cheong Leung
The University of Hong Kong, Hong Kong,
China

Synonyms

[Frequency regulation](#); [Game theory](#); [Vehicle-to-grid](#)

Definition

Vehicle-to-grid (V2G) regulation services are services in which plug-in electric vehicles, such as battery electric vehicles (BEVs), plug-in hybrid electric vehicles (PHEVs), or hydrogen fuel cell electric vehicles (FCEVs), communicate with the power grid so as to provide frequency regulation services to the grid by manipulating their charging or discharging rates.

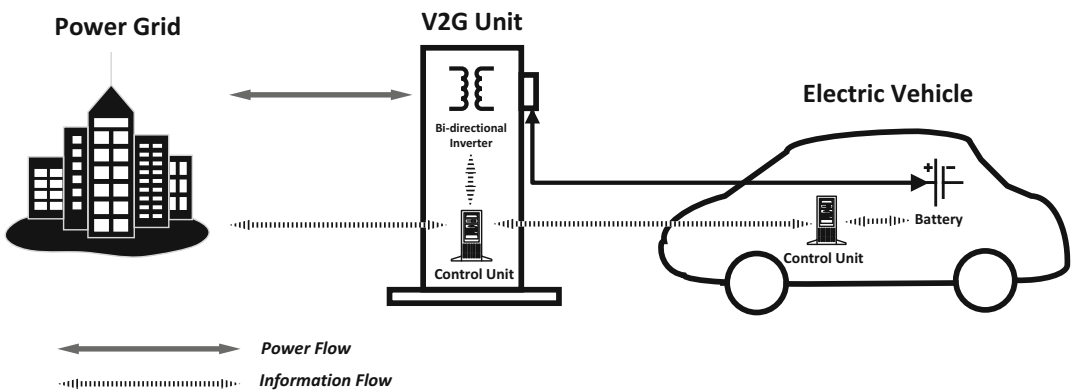
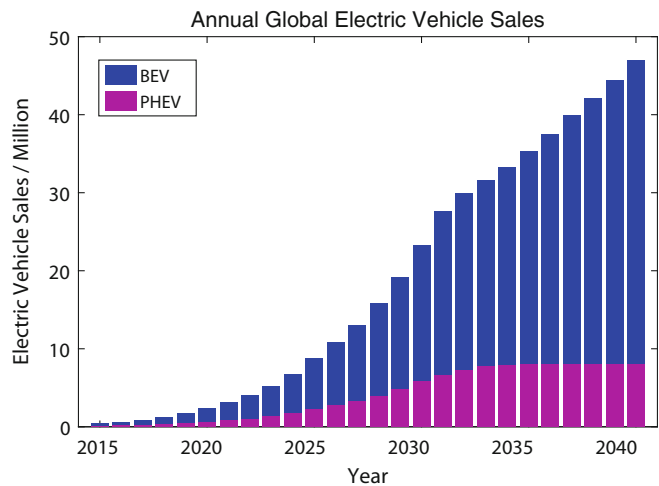
Past: Historical Background

Electric Vehicle and Vehicle-to-Grid Regulation Services

Due to the worldwide concerns on global warming, many countries have established green policies to restrain the greenhouse gas emissions, such as promoting electric vehicles (EVs). Based on the report of EV sales in 2018 (Bloomberg New Energy Finance 2018), the sales of EVs will increase from 1.1 million cars worldwide in 2017 to 11 million cars in 2025 and then toward 30 million cars in 2030 (as shown in Fig. 1) as EVs become cheaper than internal combustion engine (ICE) vehicles. The increasing penetration of EVs in transportation will not only bring environmental and economic benefits, but also help improve the reliability and stability of the power grid. Specifically, grid-connected EVs can discharge or charge their batteries to inject power to or absorb power from the grid so as to provide ancillary services to the grid. This concept is defined as the vehicle-to-grid (V2G) technique. As shown in Fig. 2, EVs charge (discharge) power from (to) the grid through a V2G unit, which includes a bidirectional inverter to enable bidirectional power exchange between EVs and the grid. With the help of smart control units, V2G unit, together with its subordinate EVs, can exchange power with the grid in response to the control signals of the grid. This enables EVs to provide ancillary services to the grid.

Among the ancillary services of the power grid, frequency regulation services are important and the most expensive ancillary services.

Game Theory Meeting
Vehicle-to-Grid
Regulation: Past,
Present, and Future,
Fig. 1 EV market trend



Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future, Fig. 2 Vehicle-to-grid diagram

Frequency regulation services aim to maintain the grid frequency at its nominal value through active power compensation, so as to keep real-time power balance in the power grid. Frequency regulation services include regulation-up, which injects power from generation assets to the grid, and regulation-down, which absorbs power from the grid to loads. Some recent studies show that the bidirectional EV chargers (Kwon and Choi 2017) and power electronics inside EVs (Merrill et al. 2015) can compensate well for frequency regulation-down and regulation-up signals through battery charging/discharging control. Therefore, an aggregation of EVs, which constitutes a distributed energy storage system as a V2G system, can bring large capacities

for providing frequency regulation services. The concept of V2G regulation services is defined to refer to the regulation services provided by a V2G system. It is able to smooth out the power fluctuations of the power grid and thus is important for supporting the integration of renewables (Kempton and Tomi 2005).

Game Theory Meets Vehicle-to-Grid Regulation Services

To provide V2G regulation services, several existing studies have considered the coordination of charging/discharging behaviors of EVs from the perspective of grid optimization. Based on the control techniques, the existing work can be divided into (1) the grid measurement

approach (Pham et al. 2016; Yang et al. 2013) and (2) the optimization-based approach (Lin et al. 2014; Shao et al. 2016; Chen et al. 2017). The former measures the grid frequency and then uses droop characteristics to compensate for the frequency deviation. It allows EVs to respond quickly to frequency regulation signals with very limited communication between EVs and the power grid (Pham et al. 2016; Yang et al. 2013). However, since this approach only relies on the frequency regulation signals which lacks sophisticated coordination scheme of EVs, suboptimal solutions are generally yielded for V2G regulation services. For the optimization-based approach, a global optimization problem is formulated to guide the charging/discharging schedules of EVs toward the global optimum. Lin et al. (2014) proposed optimal V2G control strategies for EV aggregators based on both offline scheduling and online scheduling. Decentralized algorithms were also designed to solve the optimal V2G control problem based on the gradient projection method. To coordinate a large scale of EVs, hierarchical architectures of the V2G system are well studied in recent years (Shao et al. 2016; Chen et al. 2017). A two-level charging scheduling framework was proposed for a large population of EVs (Shao et al. 2016). In this framework, the charging schedules of EVs are coordinated by local EV aggregators, and these EV aggregators are further controlled by the grid operator. The Benders decomposition approach was applied so as to solve the hierarchical charging scheduling problem. Chen et al. (2017) proposed a general framework for hierarchical V2G system, which has no restriction on the number of levels of the system. Smart V2G aggregators are employed at different levels to facilitate the grid operator coordinating the charging/discharging behaviors of EVs. Through this approach, the computational burden of the grid operator can be tremendously reduced, and the scalability of the V2G system can be enhanced.

The aforementioned approaches are from the perspective of grid optimization and assume that EVs follow the instructions of the grid operator or aggregators. Nevertheless in the real-world

operations, EVs are selfish individuals which are owned by different EV owners. These owners may be more concerned about the utilities of their EVs and the degradation issue of the EV batteries, instead of grid operations. Therefore, EVs may fail to reach an agreement with the grid because of their conflict of interests. It is thus a pressing issue on how to motivate EVs to participate into the V2G system. Game theory can provide a solution to this issue. Game theory studies how intelligent rational decision-makers act while considering the conflict and cooperation among them. Instead of following the instructions from a centralized controller, in game theory, each decision-maker makes its own actions that can maximize its utility. Applying game-theoretic approaches to V2G regulation can ensure the fulfillment of the utility of individual EV since EV can choose its optimal charging/discharging strategy to maximize its utility. Therefore, game-theoretic approaches can motivate the participation of EVs in V2G regulation. Based on the nature of the game, game-theoretic approaches can be classified as non-cooperative approach and cooperative game approach. In non-cooperative game, players are competitive and only take their own utilities as the decision goals. In contrast, in cooperative game, players collaborate with each other to some extent so as to increase their social welfare or achieve certain social goals. Some existing studies (Wu et al. 2012; Tan and Wang 2017) have already consider applying game-theoretic approaches to V2G regulation, say, using cooperative game approach (Wu et al. 2012) and non-cooperative game approach (Tan and Wang 2017). Wu et al. (2012) proposed a cooperative game-theoretic model to investigate the interaction between an aggregator and EVs in a V2G market. A backup battery bank (BBB) is also deployed in the aggregator to assure achieving the frequency regulation target. The designed interaction game is a single-level game in which EVs are players who determine their charging or discharging strategies in response to the electricity price. A decentralized coordination mechanism of EVs was designed based on game theory to achieve

the optimal performance on providing frequency regulation in a distributed fashion. When considering a large V2G system in which multiple EV aggregators provide V2G regulation services, a non-cooperative two-level game-theoretic framework was proposed (Tan and Wang 2017). In the upper level, the frequency regulation capacity bids among EV aggregators were modeled as a non-cooperative game. Based on the frequency regulation prices obtained from the non-cooperative game, in the lower level, the charging scheduling of EVs was formulated as a Markov game, in which the current move of each EV is only determined by the last moves of the other EVs instead of earlier histories of moves. It is shown that the root mean square value of frequency deviation can be reduced from 0.028 Hz to 0.009 Hz by the proposed approach.

Present: Game-Theoretic Model of Vehicle-to-Grid Regulation

Challenges

Though aforementioned work (Wu et al. 2012; Tan and Wang 2017) proposed solutions to V2G regulation using game theory, there still exist some challenges to be overcome.

- First, Tan and Wang (2017) and Wu et al. (2012) assumed that the strategy set of an individual EV only includes three states, namely, charging, idle, and discharging. This assumption oversimplifies the strategy set of EVs and thus does not consider the case that the strategy set of EVs has infinitely many elements. How to devise an infinite game in which the number of the alternatives available to each player is a continuum becomes a challenge.

- Second, Tan and Wang (2017) and Wu et al. (2012) only considered the decision-making process at each time in V2G regulation, which does not consider the time-coupling effect of the decisions. They may fail to smooth out the power fluctuations of the grid during the whole scheduling period. Therefore, how to incorporate the time-coupling effect of the decisions in the game is a challenge.

- Third, how to devise an optimal pricing scheme of the EV aggregator becomes a challenge. Tan and Wang (2017) and Wu et al. (2012) only studied the benefits of EVs and the utility grid, which fails to consider the optimal pricing scheme of the EV aggregator so as to maximize its utility.

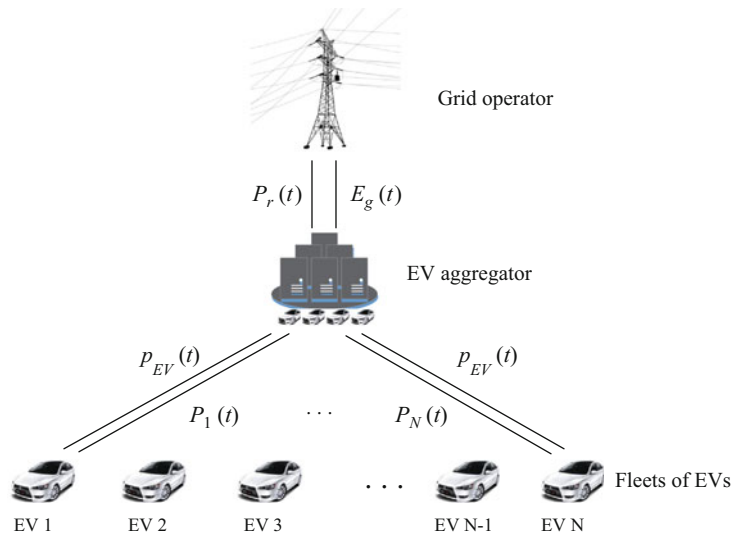
Vehicle-to-Grid Game

To deal with the aforementioned challenges, a dynamic infinite game-theoretic model of V2G regulation is introduced (Chen and Leung 2018). The designed V2G system consists of three components, namely, the grid operator, the EV aggregator, and a fleet of EVs. The V2G system aims at providing frequency regulation services to the power grid through coordinating the charging and discharging schedules of EVs. Meanwhile, EVs are required to be charged to the desired levels. As shown in Fig. 3, the EV aggregator receives regulation request $P_r(t)$ kW from the grid operator at time t and sets its electricity price as $\$p_{EV}(t)$ per kWh. EVs then determine their charging/discharging power in response to the electricity price stipulated by the EV aggregator. Note that, due to the coupling of strategies among EVs, each individual EV also considers the charging and discharging schedules of other EVs when it makes its own schedule. Based on the charging and discharging schedules of EVs, the EV aggregator buys or sells a certain amount of energy traded with EVs, namely, $E_g(t)$ kWh, from or to the grid operator, with the electricity price set in the power grid $\$p_g(t)$ per kWh.

A dynamic infinite game is considered in which the EV aggregator and EVs plan their strategies over the scheduling period $T = \{1, 2, \dots, T\}$. The number of alternatives available to each player is a continuum. Based on the nature of the game, the interaction between the EV aggregator and EVs can be formulated as a non-cooperative game or cooperative game.

In a non-cooperative game, the interaction between the EV aggregator and competitive EVs can be formulated as a Stackelberg game, which is a two-stage strategic game in which the leading decision-maker moves first and then the follow-

**Game Theory Meeting
Vehicle-to-Grid
Regulation: Past,
Present, and Future,
Fig. 3** System model



ing decision-makers move sequentially. In the upper level, EV aggregator acts as the leader who determines the optimal electricity trading price to maximize its own utility, based on the received regulation request. In the lower level, EVs are the followers who determine their charging/discharging strategies based on the price set by the EV aggregator. In order to motivate EVs to provide regulation services, incentives should be given to EVs either explicitly (through paying money for EV's regulation behavior) or implicitly (through devising pricing mechanisms so that EVs can autonomously provide regulation services while responding to the electricity price). In a cooperative game, the EV aggregator and EVs collaborate and come to an agreement to maximize their social welfare. In this sense, EV may not only concern the utility of its own but also consider the social utility (the utilities of all players) when making its strategy. The EV aggregator can give different payments to different EVs based on their behaviors in the game. Some cooperative game theories, such as Shapley value theory (Shapley et al. 1988) and potential game theory (Monderer and Shapley 1996), can be applied to analyze the interactions of players in the V2G cooperative game. Note that an underlying criteria in this cooperative game is that the utility of each player can be improved compared with the non-cooperative game. Otherwise, play-

ers will have no incentive to collaborate in this game.

Future: Research Directions

Based on the current studies on the V2G game, some possible directions are provided for future research.

Hierarchical V2G Game

In the introduced model, only the game between an EV aggregator and its subordinate EVs is considered. Nevertheless, in a large power network which includes aggregators in different levels of the system (Chen et al. 2017), hierarchical game is required to study the interaction among multiple players at different levels. Though Tan and Wang (2017) considered a two-level game in which the interactions among the grid operator, EV aggregators, and EVs were studied, the equilibrium of the game may not be reached. Therefore, it is still an open question how to validate and find the equilibrium in a general hierarchical V2G game.

Imperfect Information Game

To find the Nash equilibrium of the V2G game, EVs generally need to exchange their strategies or some information of their strategies with each

other. In this scenario, this game is a perfect information game in which players know the strategies of other players. When we consider a scenario in which EVs are not willing to share their strategies in the game with others (due to privacy concerns), this game becomes an imperfect information game. How the players find their equilibrium strategies in the imperfect information game is a promising research direction in the V2G game.

Uncertainty in the System

In the designed game, it is assumed that all the system parameters, including the regulation requests, the arrival and departure times of EVs, and the electricity price of the grid, are deterministic when players make their strategies. Nevertheless, in practical operations of the system, there exist uncertainties in these system parameters. For instance, some EVs may depart earlier than their departure times specified at their times of arrivals due to some emergencies. Therefore, how to deal with these uncertainties in the game-theoretic model becomes a challenge. Devising a robust model for the game might be an approach to overcome this challenge.

Acknowledgments This research is supported in part by the Research Grants Council, Hong Kong Special Administrative Region, China (Project No. 17261416).

References

- Bloomberg New Energy Finance (2018) Electric vehicle outlook. <http://www.cenex.co.uk/vehicle-to-grid/>
- Chen X, Leung KC (2018) A game theoretic approach to vehicle-to-grid scheduling. In: GLOBECOM 2018 – 2018 IEEE global communications conference, accepted.
- Chen X, Leung KC, Lam AYS, Hill DJ (2017) A novel online scheduling algorithm for hierarchical vehicle-to-grid system. In: GLOBECOM 2017 – 2017 IEEE global communications conference. <https://doi.org/10.1109/GLOCOM.2017.8255067>
- Kempton W, Tomi J (2005) Vehicle-to-grid power implementation: from stabilizing the grid to supporting large-scale renewable energy. *J Power Sources* 144(1):280–294. <https://doi.org/10.1016/j.jpowsour.2004.12.022>
- Kwon M, Choi S (2017) An electrolytic capacitorless bidirectional EV charger for V2G and V2H applications. *IEEE Trans Power Electron* 32(9):6792–6799. <https://doi.org/10.1109/TPEL.2016.2630711>
- Lin J, Leung KC, Li VOK (2014) Optimal scheduling with vehicle-to-grid regulation service. *IEEE Internet Things J* 1(6):556–569. <https://doi.org/10.1109/JIOT.2014.2361911>
- Merrill CH, Lam VH, Vleet MJV, Chatti MS, Brannon MC, Connelly EB, Lambert JH, Slutzky DL, Wheeler JP (2015) Modeling and simulation of fleet vehicle batteries for integrated logistics and grid services. In: 2015 systems and information engineering design symposium, pp 255–260. <https://doi.org/10.1109/SIEDS.2015.7116985>
- Monderer D, Shapley LS (1996) Potential games. *Games Econ Behav* 14(1):124–143. <https://doi.org/10.1006/game.1996.0044>
- Pham TN, Trinh H, Hien LV (2016) Load frequency control of power systems with electric vehicles and diverse transmission links using distributed functional observers. *IEEE Trans Smart Grid* 7(1):238–252. <https://doi.org/10.1109/TSG.2015.2449877>
- Shao C, Wang X, Wang X, Du C, Wang B (2016) Hierarchical charge control of large populations of EVs. *IEEE Trans Smart Grid* 7(2):1147–1155. <https://doi.org/10.1109/TSG.2015.2396952>
- Shapley L, Roth A, Press CU (1988) The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press, Cambridge
- Tan J, Wang L (2017) A game-theoretic framework for vehicle-to-grid frequency regulation considering smart charging mechanism. *IEEE Trans Smart Grid* 8(5):2358–2369. <https://doi.org/10.1109/TSG.2016.2524020>
- Wu C, Mohsenian-Rad H, Huang J (2012) Vehicle-to-aggregator interaction game. *IEEE Trans Smart Grid* 3(1):434–442. <https://doi.org/10.1109/TSG.2011.2166414>
- Yang H, Chung CY, Zhao J (2013) Application of plug-in electric vehicles to frequency regulation based on distributed signal acquisition via limited communication. *IEEE Trans Power Syst* 28(2):1017–1026. <https://doi.org/10.1109/TPWRS.2012.2209902>

Game Theory-Assisted Cooperative Spectrum Access

Jia Shi¹ and Wei Liang²

¹Xidian University, Xi'an, China

²Northwestern Polytechnical University, Xi'an, China

Synonyms

Game model-aided cooperative spectrum access

Definitions

Cooperative spectrum access technique belongs to spectrum sharing in cognitive radio networks, in which cooperative transmissions may be established between cognitive users (CUs) and primary users (PUs), or between CUs, or between PUs. The application of game theoretic including cooperative and noncooperative ones provides an enhanced spectrum utilization for cooperative spectrum access techniques.

Historical Background

In the very beginning of this century, the blending of cooperative communication technique and cognitive radio (CR) technique emerged as an important approach to improve spectrum efficiency of wireless radio resources. The radio resource is very scarce and is critical to wireless cellular networks and, thus, needs to be intelligently managed and exploited. Nevertheless, as reported in 2008, for the licensed spectrum between 30 MHz and 3 GHz in New York, the maximum occupation rate is about 13%, and the average occupation rate is around 5%. Furthermore, Zhang and Letaief (2008) introduced that the licensed spectrum reported by the Federal Communications Commission (FCC) is largely unoccupied for most of the time in the last decade, which revealed the poor exploitation of radio resources. Specifically, CR techniques allow users to share and employ the available spectrum resources which are not fully used (in time and/or spatial domain), so that the spectrum utilization can be enhanced. In a CR terminology, cognitive users (CUs) normally have the capabilities of spectrum sensing and spectrum accessing for the unoccupied radio resources of primary users (PUs). In particular, the main objective of spectrum access techniques is to enable CUs to coordinate the access with licensed spectrum of PUs. In the turn of this century, cooperative communication became an important technique for wireless systems, and it can enable

one or more nodes to relay information for a direct link by exploiting the broadcast nature of wireless channels, thereby attaining higher communication integrity and enhance channel capacity. Simeone et al. (2007) introduced that, by applying the cooperative communication techniques, it can significantly improve the performance of spectrum access in terms of the opportunity of accessing efficiency. Han et al. (2009) revealed the initial finding that the reliability of the CR networks could be largely improved by the two-hop cooperative spectrum access technique. Later on, Zou et al. (2010) studied the relay selection techniques in the context of multiple-relay CR networks with the aim of improving the performance of the second hop transmission. In recent time Zhang et al. (2014) revealed that, the main consideration in cooperative spectrum access systems was shifted to the reduction of the interference between CUs and PUs.

Foundations

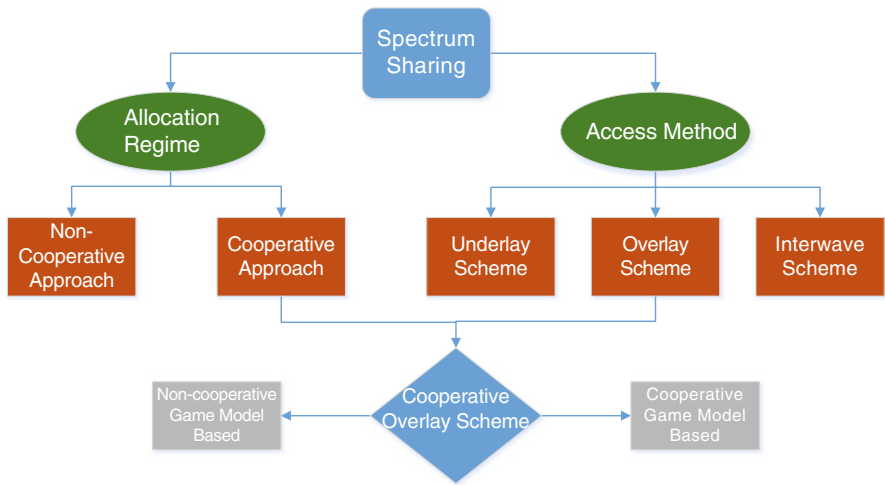
Spectrum sharing in CR terminology can be classified according to different aspects, such as in Fig. 1, including allocation regime and access method. Normally, from the perspective of spectrum allocation, there are non-cooperative and cooperative schemes. Non-cooperative regimes do not require information exchange among PUs and/or CUs, depending on how cooperation is established. As a result, it enables distributed architecture of CR network, thereby having the capability of quick response to dynamic communication environments. Nevertheless, based on limited local information, the non-cooperative approach only results in a degraded performance of spectrum efficiency. On the contrary, cooperative approaches require to share information among all users or mostly among a part of users, which need to keep the cooperation cost as low as possible for the sake of implementation complexity and energy saving. Zhao et al. (2008) investigated an approach to cluster the users and to share the information locally for implementing cooperative spectrum allocation. Compared

to non-cooperative ones, cooperative approaches exploit spectrum more efficiently, resulting in a better fairness among users.

Furthermore, shown by Fig. 1, there are three spectrum access schemes including underlay, overlay, and interwave. In the context of underlay scheme, CUs can employ all the spectrum of PUs to transmit information, subject to the condition that the interference of PUs imposed by CUs is below the tolerated threshold. Note that underlay schemes do not require to perform spectrum sensing. By contrast, for interwave schemes, CUs transmit information simultaneously with PUs only when the busy spectrum slots are detected as spectrum holes. Specifically, CUs can initiate their transmissions by exploiting the available spectrum holes not being in use for most of the time, which improves spectrum efficiency. For most overlay schemes, the cooperation between CUs and PUs can be set up, referring to as cooperative overlay spectrum access, where CUs relay PUs' information by communicating on the available spectrum simultaneously with PUs, as shown by Fig. 2. This is for CUs' payoff, and is obtained from providing the cooperative transmission for PUs, who thereby release some spectrum for CUs' transmission. Compared to underlay schemes, the interference on PUs for the overlay scheme becomes moderate due to the fact that part of the CUs' power is shifted for

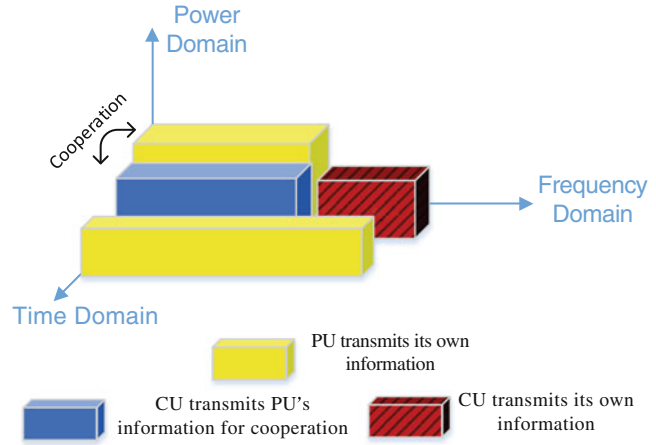
cooperative transmission. Giupponi and Ibars (2009) proposed the distributed cooperative approach for the overlay spectrum access systems, where the CUs can opportunistically access the PUs' spectrum by paying a cost for relaying PU's information. By contrast, Jiang et al. (2016) addressed the importance and challenge of knowledge sharing for cooperation between CUs and PUs in overlay spectrum access.

Haykin (2005) first introduced the application of game theory to spectrum sharing and access in CR networks, which can significantly improve spectrum efficiency of overall networks. In a game terminology, it consists of a set of players, each of which is indexed by k , a set of actions or strategies, each of which is denoted by s_k , and the payoff for each player, represented by $\pi_k(\cdot)$. In general, game theoretic techniques are classified into (1) noncooperative games, in which each player motivates to maximize its own utility by making its decision independently, and (2) cooperative games, in which by cooperation all users aim to maximize the total utility. In the majority of game models, it uses the concept of equilibrium, which ensures a player selects the optimal strategy to maximize required utility when the other players' strategies are chosen. Specifically, Han et al. (2012) introduced the definition of Nash equilibrium for CR networks: each player



Game Theory-Assisted Cooperative Spectrum Access, Fig. 1 Classification of spectrum access techniques

Game Theory-Assisted Cooperative Spectrum Access, Fig. 2 Schematic for cooperative overlay spectrum access.



cannot find a better strategy to achieve a higher payoff compared to the current one. Mathematically, the Nash equilibrium can be expressed as

$$\begin{aligned} \pi_k(s_k^*, \mathcal{S}_{K-k}^*) &\geq \pi_k(s_k, \mathcal{S}_{K-k}^*), \\ \forall k &\in \{1, 2, \dots, K\} \end{aligned} \quad (1)$$

where $\pi(\cdot)$ is the payoff conditioned on the strategies employed, s_k^* denotes the Nash equilibrium strategy for player k , and $\mathcal{S}_{K-k}^*$ is the set containing the Nash equilibrium strategies for all the players excluding play k . Note that K is the total number of users.

When applying non-cooperative game models, each user in a cooperative spectrum access system makes the selfish strategy in a distributed manner from its own perspective, which is not optimal from the holistic network-oriented perspective. The most widely used non-cooperative game models include Stackelberg game, auction game, and repeated game. Ouyang et al. (2014) proposed Stackelberg game-based waiting-time reduction approach for the overlay spectrum access systems, in which the CUs playing as relays set up cooperation with the PUs for seeking the admissions of unlicensed access in a period of time. Zhu and Hossain (2015) developed a “leader-and-follower” framework constituted by Stackelberg game model in the overlay spectrum access systems. Specifically, a PU acting as a leader can trade transmission time or power with a CU who acts as a follower for

achieving higher benefits. By contrast, Vassaki et al. (2013) applied the auction game to the overlay spectrum access in CR networks, where the auction activity happened among CUs for bidding the opportunity of relaying the PU’s information. The repeated game is usually referred to as a static noncooperative game and is also known as one-shot game. Nevertheless, the players can update their strategies according to repeating the game for multiple times; in that case, a good mutual payoff among the players may be reached. Niyato and Hossain (2008) introduced a repeated game-based mechanism for tackling the problem of spectrum pricing among PUs for offering accessing opportunities to CUs.

When a cooperative game model is employed, it usually needs a centralized unit which may be acted by one of the users for coordinating the game globally. In contrast to competing in non-cooperative games, users in cooperative games tend to cooperate with each other for the sake of maximizing total utility. In general, the spectrum access problems in CR networks were tackled by leveraging the cooperative game models including coalition games and Nash bargaining. Saad et al. (2009) started the work on using the coalition game in the spectrum access schemes, where the PUs formed a coalition for cooperation instead of competing for CUs’ access, and an increased spectrum efficiency of all PUs was derived. By contrast, El-Bessi et al. (2013) investigated the coalition game theoretic investiga-

tion of the overlay spectrum access problems, in which the CUs were formed into coalitions for cooperation with the objective of maximizing the spectrum efficiency of each coalition. Furthermore, Ma et al. (2015) proposed the coalition game-assisted spectrum access scheme in the CR networks, which motivated to maximize the utilization of available spectrum of both the PUs and CUs by formulating the cooperation among CUs. In the context of Nash bargaining games, by investigating the interaction among users, the actions of the users within a group have strong influences among themselves, so that attaining a good allocation efficiency and fairness. Attar et al. (2009) first invoked Nash bargaining game to the spectrum access mechanism for the traditional cellular networks, where the base station (BS) acting as a central unit can coordinate the spectrum and power resources for PUs and CUs while addressing the fairness issue. More recently, Zhang (2015) proposed a unified analytical framework by leveraging Nash bargaining game to solve the spectrum sharing and access problems in the cognitive small cell networks.

Key Applications

Cooperative spectrum access techniques are fundamental in various CR networks where spectrum sharing between PUs and CUs is required.

Cross-References

- [Database-Based Spectrum Access](#)
- [MAC in Cognitive Radio Networks](#)
- [Spectrum Resource Allocation in Cognitive Radio Networks](#)
- [Spectrum Sensing in Cognitive Radio Networks](#)

References

Attar A, Nakhai M, Aghvami A (2009) Cognitive radio game for secondary spectrum access problem. *IEEE Trans Wirel Commun* 8:2121–2131

- El-Bessi S, Hamouda S, Tabbane S (2013) An efficient cooperative spectrum access in cognitive radio networks via coalitional game. In: *IEEE 78th vehicular technology conference (VTC Fall)*
- Giupponi L, Ibars C (2009) Distributed cooperation in cognitive radio networks: Overlay versus underlay paradigm. In: *IEEE 69th vehicular technology conference (VTC Spring)*
- Han Y, Pandharipande A, Ting S (2009) Cooperative decode-and-forward relaying for secondary spectrum access. *IEEE Trans Wirel Commun* 8:4945–4950
- Han Z, Niyato D, Saad W, Basar T, Hjørungnes A (2012) *Game theory in wireless and communication networks: theory, models and applications*. Cambridge University Press, Cambridge
- Haykin S (2005) Cognitive radio: brain-empowered wireless communications. *IEEE J Sel Areas Commun* 23:201–220
- Jiang C, Duan L, Huang J (2016) Optimal pricing and admission control for heterogeneous secondary users. *IEEE Trans Wirel Commun* 15:5218–5230
- Ma B, Cheung M, Wong V (2015) Hybrid overlay/underlay cognitive femtocell networks: a game theoretic approach. *IEEE Trans Wirel Commun* 14:3259–3270
- Niyato D, Hossain E (2008) Competitive pricing for spectrum sharing in cognitive radio networks: dynamic game, inefficiency of Nash equilibrium, and collusion. *IEEE J Sel Areas Commun* 26:192–202
- Ouyang F, Ge J, Hou J (2014) Cooperative waiting-time reduction for cognitive radio networks using Stackelberg game. *IET Commun* 8:3072–3080
- Saad W, Han Z, Debbah M, Hjørungnes A, Basar T (2009) Coalitional game theory for communication networks. *IEEE Signal Process Mag* 26:77–97
- Simeone O, Bar-Ness Y, Spagnolini U (2007) Stable throughput of cognitive radios with and without relaying capability. *IEEE Trans Commun* 55:2351–2360
- Vassaki S, Poulakis MI, Panagopoulos AD, Constantinou P (2013) An auction-based mechanism for spectrum leasing in overlay cognitive radio networks. In: *IEEE 24th annual international symposium personal, indoor mobile radio communications (PIMRC)*, p 2733–2737
- Zhang H (2015) Resource allocation for cognitive small cell networks: a cooperative bargaining game theoretic approach. *IEEE Trans Wirel Commun* 14:3481–3493
- Zhang N, Liang H, Cheng N, Tang Y, Mark JW, Shen XS (2014) Dynamic spectrum access in multi-channel cognitive radio networks. *IEEE J Sel Areas Commun* 32:0733–8716
- Zhang W, Letaief KB (2008) Cooperative spectrum sensing with transmit and relay diversity in cognitive radio networks. *IEEE Trans Wirel Commun* 7:4761–4766
- Zhao L, Zhang J, Zhang H (2008) Using incompletely cooperative game theory in wireless mesh networks. *IEEE Netw* 22:39–44
- Zhu K, Hossain E (2015) Joint mode selection and spectrum partitioning for device-to-device communication: a dynamic Stackelberg game. *IEEE Trans Commun* 14:1406–1420

Zou Y, Zhu J, Zheng B, Yao Y (2010) An adaptive cooperation diversity scheme with best-relay selection in cognitive radio networks. *IEEE Trans Signal Process* 58:5438–5445

General Theory of Information Security

Yixian Yang
School of Cyberspace Security, Beijing
University of Posts and Telecommunications,
Beijing, China

Synonyms

[Basic theory of cyberspace security](#)

Definitions

The general theory of information security is to establish a unified basic theory of security within a certain scope.

Historical Background

In the field of information security and cyberspace security, there is no unified security basic theory. And security experts (honkers) have been working on branches of the security field to fight against hackers (Yang et al. 2016a). For instance, honkers have to study cryptographic theory for data encryption in order to prevent hackers from stealing important data; honkers have to study privacy protection technology in order to prevent hackers from stealing personal privacy through big data; honkers have to develop antivirus tools and firewalls in order to resist hackers who create destructive virus and intrude into the system illegally; honkers have to deal with all kinds of hacker's destruction activities: resist network fraud, ensure identity authentication, strengthen security management,

solve disaster recovery problems, master hacker psychology, etc.

In a word, for the areas that are closely related to people's lives, breakers (represented by the hackers) make users all over the world perturbed, especially in the field of Internet, mobile web, e-commerce, etc. so that security experts are always busy with "putting out the fire" all the time, and security field has always been in a state of "take stopgap measures (Yang et al. 2017)."

Cyberspace security has been divided into more than ten branches that look irrelevant. Cybersecurity experts have been forced to become "advanced technicians" to deal with specific problems in different branches. So that no one has the energy to consider the core issues, just like "whether there is a unified basic theory of cyberspace security" and "how to establish such a unified basic theory."

Foundations

The main research content of the general theory of information security is showed as follows: security infrastructure model in cyberspace, security meridian architecture and mechanism of cyberspace, the mesoscopic and macroscopic law of confrontation, game theory in blind and non-blind network confrontation, and the evolution of viruses, rumors, and public opinions (Yang et al. 2018).

- Security infrastructure model in cyberspace

The integrity of the system results from some synergistic effect, which generated by the interaction and interconnection of various elements within the system. Therefore, when multiple security elements and insecurity elements are mixed together, the overall influence of them will increase enormously.

In order to make the overall function of the system develop toward security, it is necessary to coordinate all elements. In the complex system of security (or insecurity), some elements are guiding and dominant, and some elements

are subordinate and dominated. In order to realize the overall security of the system, it is necessary to fully consider the relationship among various elements, strengthening the overall function of security factors and weakening the overall function of “insecurity” factors.

In view of this point, the research about security infrastructure model in cyberspace includes the following: study the “security” closely related to time, study the “security” closely related to objects in finite systems, study the “insecurity events” so as to distinguish “security event” in finite system and give the corresponding mathematical description of “security event,” study the final stability of “insecure” state in finite system based on the second law of thermodynamics, and study and solve the problem of the decomposition and prime decomposition of “insecurity event.”

- Security meridian architecture and mechanism of cyberspace

In reality, it is impossible to directly study “security” in cyberspace, which is a complex giant system (Yang et al. 2016b). Specific insecurity events can be studied to achieve the purpose of studying the “security” problem. In the cyberspace security confrontation, there are almost no concrete structures, such as parallel connection and series connection. Therefore, the “fault tree” and “event tree” in the security engineering are invalid. Traditional Chinese Medicine believes that “human body meridians” exist in the human body system. With the thoughts of Traditional Chinese Medicine, for any insecurity events in cyberspace, security completely unrelated to any specific structure can be studied, such as parallel connection, series connection, etc. With the idea of “divide and conquer,” the meridian architecture and mechanism solve the insecure problem of the small branches and then deal with the larger branches, until the overall insecurity problem is solved.

In view of this point, the research direction of cyberspace meridian architecture and

mechanism includes the decomposition of arbitrary events; the existence of “security meridian map” in limited systems and the corresponding security meridian diagram algorithm; the construction method and mechanism of “meridian map” under the “insecurity incident”; and the application range of the “meridian map” under “security events” in limited system, which includes the function of “fault tree” and “event tree.”

- Mesoscopic and macroscopic laws of confrontation

In past research of cyberspace security, security experts have focused their attentions on the local and microscopic aspects of cyber confrontation, i.e., specific core security technical issues such as encryption, virus, Trojan, and intrusion. In the environment of slow-paced confrontation, these issues are indeed the most important things in the past several decades. However, with the coming of machine hackers, the pace of “machine-to-machine offense and defense” will increase at an exponential rate. To win the initiative of cyber confrontation in the information era, we should not only understand the cyber battlefield from the microscopic view, but also from the perspective of mesoscopic and macroscopic.

In view of this point, the aspects of studying the mesoscopic and macroscopic laws of confrontation are showed as follow. With the help of dissipative structure theory, the mesoscopic description of cyberspace security confrontation can be given, so the promotion effects of the natural degradation of network systems, the attack from hackers, security measures of honkers, etc. to confrontation evolution can be analyzed, and the evaluation of the security situation of the system can be acquired; based on the general equilibrium theory of economics, the macroscopic description of network confrontation problem can be given, so the mathematical nature and operation mechanism of network confrontation problem under the condition of integrated attack and defense can be revealed, and the limitation of cyber capacity between hackers

and honkers on one-on-one confrontation and multiparty confrontation can be studied.

- Game theory in blind and non-blind cyber confrontation

Cyber confrontation is the core content of cyberspace security research field. The blind confrontation of cyber confrontation is that the both sides of attack and defense only know the result of their own assessment and know nothing about the assessment result of the opponent after the fight; the non-blind confrontation is that both sides know the consistent outcome of this round after each round of confrontation. Broadly speaking, cyber confrontation is also a special game. However, for a long time, the professional scholars haven't clearly sorted out the scenes of attack and defense in cyber confrontation and haven't explored the game theory in cyber confrontation (Lei et al. 2017).

In view of this point, the game theory in blind and non-blind confrontation can be studied from the following aspects: study the existence of Nash equilibrium between blind and non-blind confrontation processes from the perspective of game theory; study the relationship between channel capacity and mixed strategy of attack and defense corresponding to Nash equilibrium; and study the fusion of channel capacity, the core of information theory, and Nash equilibrium, the core of game theory. Try to use the multiuser advantage of game theory to make up the multiuser lack in information theory, so as to provide new ideas for basic problems of network information theory that is from information to capacity; study the mutual information between the input and output, the game theory between sender and receiver, the relationship of mutual information between the input and output, and the channel capacity between sender and receiver while the channel is settled.

- The evolution of viruses, rumors, and public opinions

In the cyber war of the information era, three special security attack and defense problems, computer viruses, Internet rumors, and public opinions, will attract more and

more academic attention. Computer viruses can be said to be one of the most core security issues in cyberspace, all kinds of malicious codes just like biological viruses; they cannot be exterminated and cost a large amount of resources; however, Internet rumors are regarded as one of the most lethal weapons in future information warfare by cybersecurity experts; moreover, there is no previous experience to learn and we must rely on our own efforts. Election-related public opinion is a new weapon in the cyber warfare; with the rapid development of "we-media," it will be more and more lethal to a certain kind of country. Therefore, these security methods must be systematically studied, and their corresponding research methods can be extended to other security fields.

In view of this point, the evolution of viruses, rumors, and public opinions can be studied from several aspects: study the transmission dynamics of several types of viruses (e.g., death virus, rehabilitation virus, and immune virus); study the effects of several prevention measures, analyze the situation of latent virus, and explore virus mechanisms and prevention mechanisms; study the dynamics of rumor under the situations of single institution and multiple institutions; study the dynamic equations of public opinion structure and the steady-state solutions of several kinds of public opinion equations; and reveal the evolution rules of viruses, rumors, and public opinions.

Key Applications

The general theory of information security tries to stipulate corresponding security implications and establish a unified security foundation theory within a certain scope. It makes a quantitative discussion on security confrontation in cyberspace, comprehensively and systematically. The general theory of information security aims to reveal some basic rules of hackers' attack and defense and security evolution and applies them to the main branches of cyberspace security.

Cross-References

- [Access Control](#)
- [Mobile Security and Privacy](#)
- [Network Privacy Surveillance and Privacy Protection](#)

References

- Lei M, Yang Y, Niu X et al (2017) An overview of general theory of security. *China Commun* 14(7):1–10
- Yang Y, Niu X, Li L et al (2016a) General theory of security and a study of hacker's behavior in big data era. *Peer-to-Peer Netw Appl* 11(1):1–10
- Yang Y, Zhang Z, Li W (2016b) Intelligent information network security and management. *China Commun* 13(7):iii–vi
- Yang Y, Peng H, Li L et al (2017) General theory of security and a study case in Internet of Things. *IEEE Internet Things J* 4(2):592–600
- Yang Y, Niu X, Li L et al (2018) A secure and efficient transmission method in connected vehicular cloud computing. *IEEE Netw* 32(3):14–19

Geometrical Probability

- [Random Distance Distribution](#)

Ginibre Point Process

- [Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes](#)

Green C-RAN with Network Slicing

- [Network Slicing-Enabled Green C-RAN](#)

Green Networking

- [Energy Efficiency of Cellular Networks](#)

Green Radio

- [Energy Efficiency of Cellular Networks](#)

Green Wireless Communications

- [Energy Efficiency of Cellular Networks](#)

Grid Monitoring and Anomaly Detection Using Power Line Communications

- [Power Line Communications for Grid Discovery and Diagnostics](#)

Grid Surveillance

- [Power Line Communications for Grid Discovery and Diagnostics](#)

Group Authentication

- [Wireless Group Authentication](#)

Group Key Agreement for Wireless Networks

Tianqi Zhou
School of Computer and Software, Nanjing
University of Information Science and
Technology, Nanjing, China

Synonyms

[Multiparty key agreement for wireless networks](#)

Definitions

Group key agreement is the protocol whereby multiple parties can agree on a common key, which ensures secure information transmission in their later communication. That is, the negotiated common key can be used to encrypt message for communicating parties. Protocols that are useful in practice do not reveal the negotiated common key to any eavesdropping parties.

In wireless networks, group key agreement refers to a group of m trusted nodes that aim to agree on a common secret key K over a wireless channel. Note that the communication channel for agreeing on the common secret key K is insecure. That is, the communication channel can be controlled by an adversary. In general, there are two types of adversaries: passive adversary and active adversary.

1. **Passive adversary:** A passive adversary tries to learn information about the common secret key by eavesdropping on the communication channel.
2. **Active adversary:** An active adversary attempts to impersonate other party or disrupt the protocol by altering, injecting, intercepting, and replaying messages.

The goal of group key agreement is to agree on a common secret key without revealing the common secret key to any adversaries.

Historical Background

Diffie-Hellman protocol (Diffie and Hellman 1976) is the first modern key agreement protocol, which is proposed in 1976. But this protocol can only support two parties. By using Weil pairing, Joux proposed the three-party key agreement protocol (Joux 2000). Note that this protocol requires only one round to derive the common secret key among three parties.

In order to generate the common secret key among multiple parties, Steiner et al. proposed the key agreement protocol (Steiner et al. 2000), which supports multiple parties and

considers dynamic change among these parties. However, this protocol needs $n + 1$ rounds to generate the common secret key. In addition, the communication and computation overhead are relatively large. Here, n is the number of parties. Based on Joux's protocol, Barua et al. proposed the multiparty key agreement protocol (Barua et al. 2003). In this protocol, a tree structure is employed to help multiple parties to agree on a common secret key. Note that for n parties, it requires $n \lceil \log_3 n \rceil$ rounds to derive the common secret key. And the communication and computation overhead are reduced compared with Steiner et al. (2000). Later, Shen et al. proposed the group key agreement protocol (Shen et al. 2018) that aims to reduce the communication overhead and the number of rounds for generating the common secret key. In particular, the communication overhead and the rounds of the protocol are $n\sqrt{n}$ and 2, respectively.

The above group key agreement protocols are interactive key agreement protocols. Namely, each party requires to interact with other parties to derive a common secret key. Compared with interactive key agreement protocol, noninteractive key agreement protocol is the other kind of key agreement protocol, where multiple parties can agree on a common secret key without interacting with others. In 2003, by using the concept of multilinear maps, Boneh and Silverberg presented a multiparty noninteractive key agreement protocol (Boneh and Silverberg 2003), whereas it was not until 2013 that the constructions for multilinear map was designed by Garg, Gentry, and Halevi using ideal lattices (Garg et al. 2013). Meanwhile, Coron, Lepoint, and Tibouchi also presented the constructions for multilinear map over integers (Coron et al. 2013) at the same year. However, it is pointed in Boneh and Zhandry (2017) that both Garg et al. (2013) and Coron et al. (2013) require a secure setup for key agreement purpose. By using the concept of indistinguishability obfuscation and the constrained pseudorandom functions, Boneh and Zhandry designed the noninteractive key agreement protocol that requires no trusted setup.

Foundations

Security assumptions and adversarial model are two essential foundations for group key agreement in wireless networks. For one thing, the security of the protocol can be deduced from security assumptions. For another thing, the security level of the protocol depends on the strength of the adversarial model that is defined previously. In the following, some security assumptions and a well-known adversarial model of group key agreement are presented.

1. Security assumptions.

- (a) Decisional Diffie-Hellman (DDH) assumption: For a group G of order q , the DDH assumption says that the following two tuples is computationally indistinguishable.

$$\begin{cases} R = (g, g^x, g^y, g^z) \in G^4 \\ D = (g, g^x, g^y, g^{xy}) \in G^4 \end{cases}$$

Here, g is the generator of G , and $x, y, z \leftarrow_R \mathbb{Z}_q$.

- (b) Computational Diffie-Hellman (CDH) assumption: For a group G of order q , the CDH assumption states that given

$$g, g^x, g^y$$

it is hard to compute

$$g^{xy}$$

Here, g is the generator of G , and $x, y \leftarrow_R \mathbb{Z}_q$.

- (c) Bilinear Diffie-Hellman (BDH) assumption: For groups G_1 and G_2 of order q , the BDH assumption states that given

$$g, xg, yg, zg$$

it is hard to compute

$$\hat{e}(g, g)^{xyz}$$

Here, g is the generator of G_1 , and $x, y, z \leftarrow_R \mathbb{Z}_q$.

2. Adversarial model. The capabilities and possible actions of the attacker can be reflected by the adversarial model. One of the famous adversarial models is the unauthenticated links model (Canetti and Krawczyk 2001), which is described as follows.

- (a) Basic capabilities: Basic capabilities of the attacker include listening to all the transmitted information, controlling the transmission way of the messages, changing messages, or injecting messages.
- (b) Obtaining secret information: Attacker can access secret information of parties. According to the type of information that can be accessed by the attacker, three categories of information leakage are summarized as follows.
- (i) Session-state reveal.
 - (ii) Session-key query.
 - (iii) Party corruption.

In addition to the above capabilities provided for the attacker, session expiration is also an important element of the adversarial model. That is, the value of the expired session key can no longer be revealed by the attackers. Note that it is the foundation for the notion of perfect forward secrecy in key agreement.

Applications

The group key agreement is used to establish a common secret key such that the communications among multiple parties encrypted with this key are protected from the attacks of adversaries. In particular, the generated common secret key can be used for encrypting messages among multiple nodes in wireless networks. Note that the common secret key hold by multiple nodes supports the symmetric encryption for the message, which is more efficient than asymmetric encryption and has wide applications in wireless networks.

Cross-References

- [Access Control](#)
- [Authentication](#)
- [Mobile Security and Privacy](#)
- [Privacy-Preserving Data Collection](#)

References

- Barua R, Dutta R, Sarkar P (2003) Extending joux protocol to multi party key agreement. In: International conference on cryptology in India. Springer, pp 205–217
- Boneh D, Silverberg A (2003) Applications of multilinear forms to cryptography. *Contemp Math* 324(1):71–90
- Boneh D, Zhandry M (2017) Multiparty key exchange, efficient traitor tracing, and more from indistinguishability obfuscation. *Algorithmica* 79(4):1233–1285
- Canetti R, Krawczyk H (2001) Analysis of key-exchange protocols and their use for building secure channels. In: International conference on the theory and applications of cryptographic techniques. Springer, pp 453–474
- Coron JS, Lepoint T, Tibouchi M (2013) Practical multilinear maps over the integers. In: Advances in cryptology—CRYPTO 2013. Springer, pp 476–493
- Diffie W, Hellman M (1976) New directions in cryptography. *IEEE Trans Inf Theory* 22(6):644–654
- Garg S, Gentry C, Halevi S (2013) Candidate multilinear maps from ideal lattices. In: Annual international conference on the theory and applications of cryptographic techniques. Springer, pp 1–17
- Joux A (2000) A one round protocol for tripartite Diffie–Hellman. In: International algorithmic number theory symposium. Springer, pp 385–393
- Shen J, Zhou T, Chen X, Li J, Susilo W (2018) Anonymous and traceable group data sharing in cloud computing. *IEEE Trans Inf Forensics Secur* 13(4):912–925
- Steiner M, Tsudik G, Waidner M (2000) Key agreement in dynamic peer groups. *IEEE Trans Parallel Distrib Syst* 11(8):769–780

Handoff in LTE-U/LAA

► [Handover in LTE-U/LAA](#)

Handoff Management in Wireless Communication-Based Train Control Systems

Li Zhu¹ and Richard Fei Yu^{2,3}

¹Beijing Jiaotong University, Beijing, China

²Carleton University, Ottawa, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

[CBTC](#); [SCTP](#); [SMDP](#)

Definitions

Communication-based train control (CBTC) system is an automated control system for railways that ensures the safe operation of rail vehicles using wireless data communications. It can maximize the utilization of railway network infrastructure and enhance the level of safety and service offered to customers. The train-ground

communication system is one of the key sub-systems of CBTC. A seamless handoff scheme based on SCTP and IEEE 802.11p is proposed to provide high link availability in CBTC systems. Stochastic semi-Markov decision process (SMDP) has been successfully used to solve finance and admission control problems, among others. This article focuses on the application of SMDP to the handoff decision problem in CBTC systems. Maximizing the SCTP throughput and minimizing the handoff latency are the objectives in the handoff decision phase.

Historical Background

Communications-based train control (CBTC) systems is an automated train control system using continuous, high capacity, and bidirectional train-ground communications to ensure the safe operation of rail vehicles (Pascoe and Eichorn 2009). It can maximize the utilization of railway network infrastructure and enhance the level of safety and service offered to customers. As numerous CBTC systems are in operation around the world, CBTC is becoming more and more popular as an advanced train control system.

CBTC systems have stringent requirements for wireless communication availability and latency. Whereas in commercial wireless networks, less service availability means less revenues in CBTC systems, less service availability could cause train derailment, collision, or even catastrophic loss of life

or assets. Therefore, although it may not be necessary to have continuous communications between the train and ground, it is important that the radio links are available when they are needed and the latency is minimized. Furthermore, it is important to maintain this availability and quick response time, while a train is moving at high speed.

Most existing WLAN-based train-ground systems are using traditional IEEE 802.11 technologies, such as 802.11a/b/g. However, IEEE 802.11a/b/g WLANs were not originally designed for high speed environments. Particularly, when a train moves away from the coverage of a WLAN access point (AP) and enters the coverage of another AP along the railway, a handoff procedure occurs, and this handoff process may result in communication interrupt and long latency (Mishra et al. 2003).

There are numerous handoff schemes proposed in the literature to decrease WLAN handoff latency (Mishra et al. 2003; Ramani and Ramani 2005) in data link layer, but most of them require coordination/cooperation between APs and mobile clients, which may make it difficult to implement them in CBTC systems.

The seamless handoff management problem can also be solved at transport layer. Stream Control Transmission Protocol (SCTP) (Ma et al. 2004), a new IETF-standardized transport layer protocol in addition to Transmission Control Protocol (TCP) and User Datagram Protocol (UDP), is commonly used to solve the handoff management problem. As a protocol with the features of the multi-homing, multi-stream, and partial reliable data transmission, SCTP are especially attractive for applications that have stringent performance and high reliability requirements. Compared to the data-link-layer handoff management approaches, transport layer schemes have many advantages, such as improved throughput and latency performance. Moreover, no third party other than the endpoints participates in handoff process, and no modification or addition of network components is required, which makes transport layer approaches attractive in WLAN-based CBTC systems, where commercial-off-the-shelf equipments are widely used.

Although some works have been done for handoff management, most of the works are about general commercial wireless system, which is different from the CBTC train-ground communication system.

In this article, we propose a seamless handoff scheme for CBTC train-ground communication system based on SCTP and IEEE 802.11p, which is an emerging technology for vehicular communication networks. To the best of our knowledge, the design of seamless handoff scheme based on SCTP and IEEE 802.11p has not been done in previous works. The distinct features of the proposed scheme are as follows.

- (1) We propose a seamless handoff scheme based on SCTP and IEEE 802.11p to provide high link availability in train-ground communication systems.
- (2) The handoff decision problem is formulated as a stochastic semi-Markov decision process (SMDP), which has been successfully used to solve finance (Feinberg and Schwartz 2002) and admission control (Yu et al. 2006) problems, among others.
- (3) The simulation results show that the proposed scheme can significantly improve SCTP throughput and achieve seamless handoff in train-ground communication system.

Proposed Train-Ground Communication System Based on SCTP and IEEE 802.11p

Overview of Communication-Based Train Control and Its Train-Ground Communication System

Figure 1 gives a simple view of a CBTC system with WLAN-based train-ground communication system. In this system, continuous bidirectional wireless communications between each station adapter (SA) on the train and the wayside access point (AP) are adopted instead of the traditional fixed-block track circuit. The railway line is usually divided into areas. Each area is under the control of a zone controller (ZC) and has its

own wireless transmission system. The identity, location, direction, and speed of each train are transmitted to the ZC. The wireless link between each train and the ZC must be continuous to ensure the ZC knows the location of all the trains in its area at all times. The ZC transmits to each train the location of the train in front of it; a braking curve is given to enable it to stop before it reaches that train as well. Theoretically, two successive trains can travel together as close as a few meters in between them, as long as they are traveling at the same speed and have the same braking capability.

Train-ground communication systems are primarily designed to connect each component of CBTC systems: zone centers (ZCs), access points (APs) along a railway, and train aboard equipment. A basic configuration of a WLAN-based train-ground communication system is shown in Fig. 1. Following the philosophy of open standards and interoperability, the backbone network of the train-ground communication system includes Ethernet switches and fiber-optic cabling that are based on the IEEE 802.3

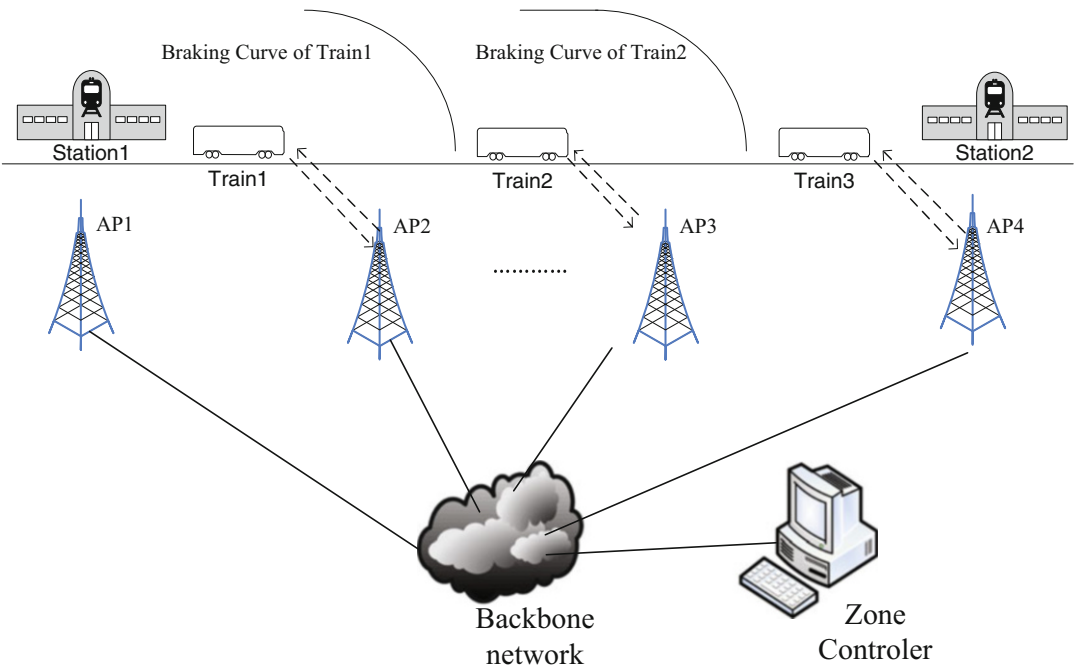
standard. The wireless portions of the train-ground communication system, which consist of APs along the railway and SAs on the train, are based on the IEEE 802.11 standard.

When a train moves away from the coverage of an AP and enters the coverage of another AP along the railway, the handoff procedure may result in communication interrupt and long latency. In CBTC systems, it is important to maintain communication link availability in order to guarantee train operation safety and efficiency. To this end, we present a seamless handoff scheme based on SCTP at transport layer and IEEE 802.11p to provide high link availability in CBTC systems.

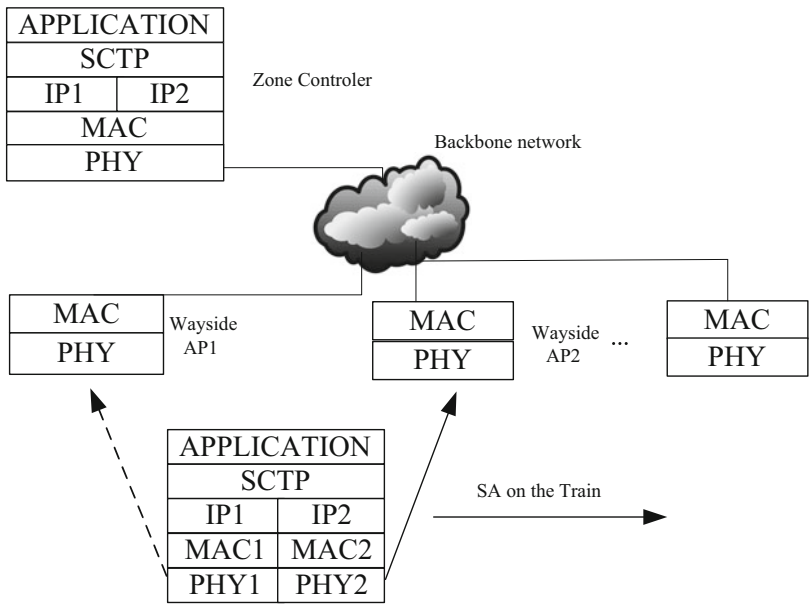
The Proposed Train-Ground Communication System Based on SCTP and IEEE 802.11p

Figure 2 describes the system architecture and protocol stack of the proposed train-ground system based on SCTP and IEEE 802.11p. There are two radios in the SA on the train. Each radio

H



Handoff Management in Wireless Communication-Based Train Control Systems, Fig. 1 A communication-based train control (CBTC) system



Handoff Management in Wireless Communication-Based Train Control Systems, Fig. 2 System architecture and protocol stack of the proposed train-ground communication system based on SCTP and IEEE 802.11p

is related to a different MAC interface and IP interface. The ZC also has two IP interfaces.

Normally only one pair of IP addresses (the IP address in the ZC and the IP address in the SA) is working, which is called the primarily path. When a handoff is triggered, the SA on the train will try to exchange necessary information with the ZC to establish another path with the other radio, which is called the secondary path. When the second path is established, the communication nodes copy all the buffered data from the primary path to the second path so that the data in sending buffers of the two paths are completely the same. The SA also needs to cut off the primary path to finish the handoff process at an appropriate time. The whole handoff procedure is shown in Fig. 3.

A critical issue in the above train-ground communication system is the handoff decision policy, i.e., when to perform handoff. In high speed environments, wireless channels are changing dynamically in CBTC systems. If the handoff decision policy is not designed carefully, communication interrupt and long latency may occur, which will significantly affect

the performance of a CBTC system. Therefore, an efficient handoff decision policy is needed to decide at what time the second path should be established and when to cut off the primary path, which will be studied in the following section.

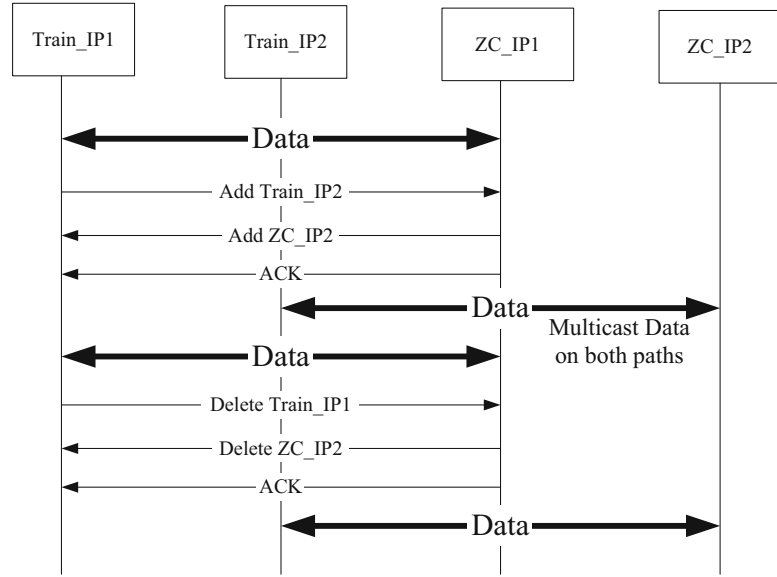
Optimal Handoff Decision Policy in the Train-Ground Communication System Based on SCTP and IEEE 802.11p

An SMDP model consists of the following five elements: (1) decision epochs, (2) states, (3) actions, (4) rewards, and (5) transition probabilities, which will be described in the following.

Action and State

In our SMDP model, at each decision epoch, the SA on the train has to decide whether the connection should use the current chosen AP or connect to the next AP (we assume the SA on the train won't be in the coverage of three successive APs).

Handoff Management in Wireless Communication-Based Train Control Systems, Fig. 3 The proposed handoff procedure



Each state has the following information:

1. The measured signal noise ratio from two APs;
2. The current SCTP congestion window;
3. The path(s) currently used by the SA;

Reward Function

In our formulation, we define the reward function as the combination of SCTP throughput and delay. This is a common approach used in the optimization literature, which is called Aggregate Objective Function (AOF), to solve an optimization problem with multiple objectives. In reality, different CBTC networks have different throughput and packet delay requirements. By adjusting the parameters in the reward function, the proposed scheme is generic enough to accommodate different requirements in real CBTC networks.

State Transition Probability

In CBTC systems, due to the restrictions of railways and trains, the ground antenna cannot be installed too high. The low antenna height means that the Fresnel zone typically limits the propagation of radio to fairly short ranges.

In this paper, we use finite-state Markov channel (FSMC) models in CBTC systems.

FSMC models have been widely accepted in the literature as an effective approach to characterize the correlation structure of the fading process, including satellite channels, indoor channels (Babich and Lombardi 1997), Rayleigh fading channels (Wang and Chang 1996), and Nakagami fading channels (Iskander and Mathiopoulos 2003). Considering FSMC models may enable substantial performance improvement over the schemes with memoryless channel models (Yang et al. 2005).

In FSMC models, the range of the received SNR can be partitioned into discrete levels. Each level corresponds to a state in the Markov chain. Assume there are L states in the model. Let i and γ denote the instantaneous channel state and SNR, respectively. When the channel is in state i , the corresponding SNR is γ_i . Then we have $\gamma_i < \gamma < \gamma_{i+1}$, $1 \leq i \leq L - 1$. The probability of transition from state i to state j in the Markov model is denoted by $P_{i,j}$.

In real systems, the values of the above transition probability can be obtained from the history observation of CBTC systems.

In order to derive the congestion window transition probability, we refer to the SCTP behavior model in Ma et al. (2007). The SCTP behavior is modeled in terms of "rounds," where a round starts when the sender begins the transmission of

a window of chunks and ends when the sender receives the last acknowledgment for chunks in this window. SCTP essentially doubles its congestion window size in the slow-start stage and increases the congestion window linearly in the congestion avoidance stage if none of the chunks in the previous window is lost during the previous RTT . If one or more chunks in the previous window are lost, the congestion window is set to half of the current window. The congestion window will not change until all the chunks in the window are sent out.

Given the current congestion window η , when SCTP is in slow-start, the congestion window in the next epoch can be η , $\eta/2$, or 2η . For the congestion avoidance stage, the congestion window can be changed to η , $\eta/2$, or $\eta + 1$, which depends on the decision epoch and packet losses.

Optimality Equations and Value Iteration Algorithm

The optimality equation for an SMDP is given by Puterman (1994):

$$v(s) = \max_{a \in A} \left\{ r(s, a) + \sum_{s' \in S} \lambda P[s'|s, a] v(s') \right\}, \quad (1)$$

where $v(s)$ denotes the maximum expected total reward, given the initial state s , and s' represents next state. That is

$$v(s) = \max_{\pi} v^{\pi}(s). \quad (2)$$

The solutions of the optimality equation correspond to the maximum expected total reward $v(s)$ and the SMDP optimal policy δ^* . Note that the SMDP optimal policy δ^* indicates the decision as to which action to choose.

We use value iteration algorithm (VIA) in this paper to determine a stationary deterministic optimal policy and the corresponding expected total reward.

Simulation Results and Discussions

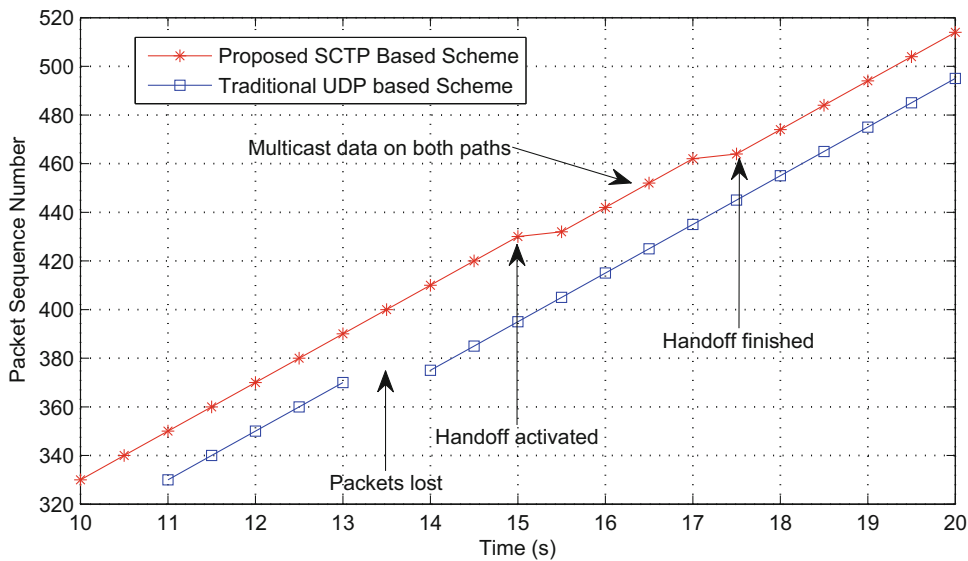
In this section, simulation examples are used to illustrate the performance of the proposed handoff scheme based on SCTP and IEEE 802.11p. We compare the handoff delay in the proposed scheme with that of existing handoff scheme based on UDP and traditional IEEE 802.11. The results show that the handoff delay is very close to zero in the proposed scheme. In addition, the simulation results show that the proposed SMDP based optimal handoff policy can significantly improve SCTP throughput in CBTC train-ground communication systems.

Figure 4 shows the sequence numbers of the packets at the corresponding time on the X axis. From this figure, we can see that, in the traditional UDP and 802.11 based handoff scheme, the SA on the train cannot have communication with any of AP in a period of time, and the packet sent in that period of time is lost which is unacceptable for real-time and safe required applications. This is caused by the handoff delay in the traditional handoff scheme, in which the old path is broken before the new path is setup to the new AP.

Comparatively, it is evident that the proposed handoff scheme supports seamless handoff between adjacent APs. During the handoff process, the SA associates to the new AP with the other radio, obtains a new IP address from the old path, and communicates with the new address before the old path is terminated. No packets are lost during this seamless handoff procedure. The slight delay in the handoff process is caused by the information exchange before the establishment of the new path.

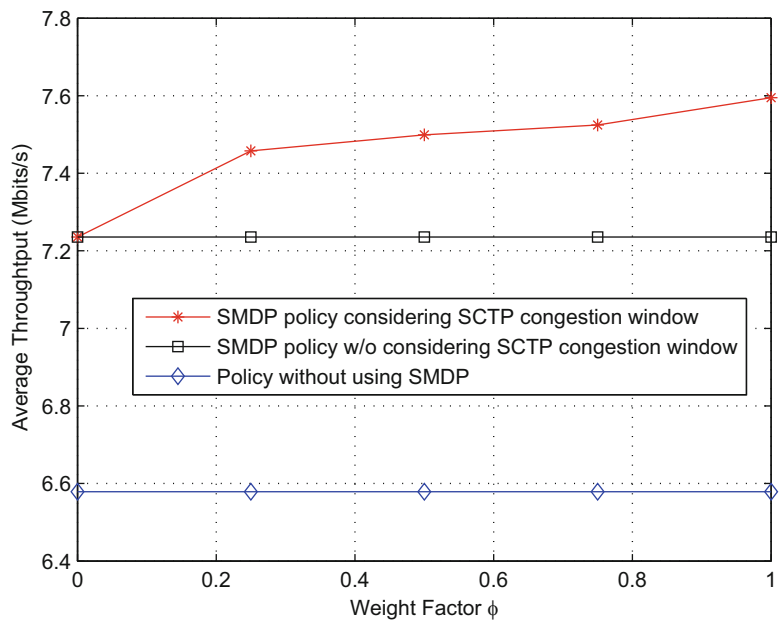
We compare the performance of our proposed handoff decision policy, with the other two heuristic handoff decision policies. For the first heuristic policy, the path to be selected in each decision epoch is the one that always has the better SNR. For the second heuristic policy, we also model the handoff decision as a SMDP, but SCTP congestion window variation is not considered in this model.

The SCTP throughput improvement is shown in Fig. 5. The SCTP throughput increases with the weight factor for our proposed



Handoff Management in Wireless Communication-Based Train Control Systems, Fig. 4 Packet sequence number in different schemes

Handoff Management in Wireless Communication-Based Train Control Systems, Fig. 5 The throughput under three policies with different weight factor



policy; that is because the policy is designed to get more throughput when weight factor increases. Similarly, the two heuristic policies give unchanged worse throughput performance compared to our proposed decision policy when weight factor increases; this is because these two policies don't care too much about

SCTP congestion window variation which is a very important information for throughput. Particularly, when the weight factor decreases to zero, our proposed decision policy is not designed to get the best SCTP throughput, which gives the same throughput performance as one of the heuristic policies.

References

- Babich F, Lombardi G (1997) A measurement based Markov model for the indoor propagation channel. In: *Proceedings of IEEE 47th VTC, Phoenix*, vol 1, pp 77–81
- Feinberg EA, Schwartz A (eds) (2002) Markov decision processes in finance and dynamic options. In: Feinberg EA, Schwartz A (eds) *Handbook of Markov decision processes*. Kluwer, Boston
- Iskander CD, Mathiopoulos PT (2003) Fast simulation of diversity Nakagami fading channels using finite-state Markov models. *IEEE Trans Broadcast* 49(3): 269–277
- Ma L, Yu F, Leung VCM (2004) A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wirel Commun* 11(4):44–51
- Ma L, Yu F, Leung VCM (2007) Performance improvements of mobile SCTP in integrated heterogeneous wireless networks. *IEEE Trans Wirel Commun* 6:3567–3578
- Mishra A, Shin M, Arbaugh W (2003) An empirical analysis of the IEEE 802.11 MAC layer handoff process. *SIGCOMM Comput Commun Rev* 33:93–102
- Pascoe RD, Eichorn TN (2009) What is communication-based train control. *IEEE Veh Technol Mag* 4:16–21
- Puterman M (1994) *Markov decision processes: discrete stochastic dynamic programming*. Wiley, New York
- Ramani I, Ramani I (2005) Syncscan: practical fast handoff for 802.11 infrastructure networks. In: *Proceedings of IEEE Infocom'05, Seoul*, pp 675–684
- Wang HS, Chang PC (1996) On verifying the first-order Markovian assumption for a rayleigh fading channel model. *IEEE Trans Veh Technol* 45(2):353–357
- Yang J, Khandani AK, Tin N (2005) Statistical decision making in adaptive modulation and coding for 3G wireless systems. *IEEE Trans Veh Technol* 54(6): 2066–2073
- Yu F, Krishnamurthy V, Leung VCM (2006) Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless CDMA networks. *IEEE Trans Signal Process* 54(2):542–555

Handoffs, Modeling in IP Networks

Jong-Hyoun Lee

Department of Software, Sangmyung University,
Cheonan, Republic of Korea

Synonyms

[Handover](#)

Definition

A handoff in IP networks is the process by which a mobile node changes its point of attachment from one network to another network while maintaining IP communication sessions (Lee et al. 2013a). Without the IP handoff support for the mobile node, all ongoing IP communication sessions are disrupted by the change of point of attachment. To support a seamless handoff in IP networks, specially designed network entities and protocols are required, and those have been developed for IPv4 networks and IPv6 networks separately.

Historical Background

As the handoff of a mobile node in IP networks is related to network entities and protocols for the Internet, the Internet Engineering Task Force (IETF) are taking the lead to development of specifications for the seamless handoff. The IP handoff support for IPv4 called Mobile IPv4 (MIPv4) was first specified in RFC 3344 and obsoleted by RFC 5944. The IP handoff support for IPv6 called Mobile IPv6 (MIPv6) was first specified in RFC 3775 and obsoleted by RFC 6275. Such host-based mobility management protocols (i.e., MIPv4 and MIPv6) assume that a mobile node is capable to signal to the network entities that the mobile node changes its point of attachment from one network to another network. This assumption does not hold for all cases especially for lightweight mobile devices such as small sensors and mobile phones. Proxy Mobile IPv6 (PMIPv6) specified in RFC 5213 was developed to solve this issue as a network-based mobility management protocol. Network entities detect the IP handoff of a mobile node and provide required operations for IP session continuity in PMIPv6.

The IP handoff directly affects the user quality of experience (Pack et al. 2007). A long handoff latency results in a high packet loss and thus causes poor quality of experience. In addition, depending on a mobility management

protocol, the required operational cost for the IP handoff support is different. Accordingly, various extensions of the mobility management protocols have been specified by the IETF, e.g., fast handover, route optimization, and hierarchical mobility management extensions. Also, various improvements have been suggested from researchers over years.

Key Points

The IP handoff is commonly evaluated with the following performance criteria:

- Handoff latency: It is the time interval during which a mobile node cannot send or receive any packets, while it changes its point of attachment from one network to another network (Lee et al. 2013a).
- Handoff blocking probability: It is the probability which a mobile node cannot complete its handoff when the network residence time is less than the handoff latency (Lee et al. 2013a).
- Packet loss: It is the sum of all lost packets destined for a mobile node during the mobile node's handoff (Lee et al. 2013a).
- Signaling cost: It is the accumulative signaling overhead for the IP handoff support (Lee et al. 2010).
- Packet delivery cost: It is the accumulative traffic overhead caused by packet delivery on the routing path for the IP session continuity (Lee et al. 2010).

For calculating the handoff latency, a timing diagram of the handoff is required, and owing to the operational differences, each mobility management protocol presents different timing diagrams for the IP handoff support. For instance, the handoff latency for MIPv6 is the sum of the link-layer handoff latency, the movement detection latency at the network layer, the duplicate address detection latency, and the location registration latency, whereas the handoff latency for PMIPv6 is the sum of the link-layer hand-

off latency and the location registration latency. Also, fast handover extensions further minimize the handoff latency by performing IP handoff required operations in advance.

The handoff blocking probability and the packet loss are proportional to the handoff latency as expected. The higher handoff latency is essentially a trouble resulting in poor performance which must lead to poor quality of experience for users.

The signaling cost is calculated as the product of the size of signaling message for the IP handoff support and the hop distance between the signaling message source and destination. The required signaling messages and the hop distance are different with mobility management protocols. As extensions are applied to the basic mobility management protocols like MIPv6 and PMIPv6, the protocol complexity increases and the signaling cost thus increases.

The packet delivery cost is calculated as the product of the data packet size and the hop distance between the data packet source and destination. The hop distance is different with mobility management protocols.

Key Applications

The understandings of the IP handoff support and related performance criteria are important when we choose an appropriate IP mobility management protocol for a given condition. Those are also required when we design new IP mobility management protocols.

Future Directions

The initial approaches for the IP handoff support such as MIPv4, MIPv6, and PMIPv6 were developed based on centralized mobility anchors that facilitate the IP handoff support for mobile nodes in a highly efficient manner. When those mobility management protocols were developed, mobile network architectures were in centralized hierarchical forms. The deployment of such centralized mobility anchors in the centralized

hierarchical forms was thus considered. However, due to rapidly increasing Internet traffic, the mobile network architectures are being changed toward flat mobile network architectures in 5G. The new approach called distributed mobility management is thus being considered for the flat mobile network architectures in 5G (Lee et al. 2013b). The IP handoff support is yet the main key for distributed mobility management (Ko et al. 2016).

Cross-References

- [Distributed IP Mobility Management](#)
- [Mobile IP](#)
- [Network-layer Mobility Management](#)
- [Proxy Mobile IPv6](#)

References

- Ko H, Lee G, Suh D, Pack S, Sherman Shen X (2016) An optimized and distributed data packet forwarding scheme in LTE/LTE-A networks. *IEEE Trans Veh Technol* 65(5):3462–3473
- Lee J-H, Ernst T, Chung T-M (2010) Cost analysis of IP mobility management protocols for consumer mobile devices. *IEEE Trans Consum Electron* 56(2):1010–1017
- Lee J-H, Bonnin J-M, You I, Chung T-M (2013a) Comparative handover performance analysis of IPv6 mobility management protocols. *IEEE Trans Ind Electron* 60(3):1077–1088
- Lee J-H, Bonnin J-M, Seite P, Anthony Chan H (2013b) Distributed IP mobility management from the perspective of the IETF: motivations, requirements, approaches, comparison, and challenges. *IEEE Wirel Commun Mag* 20(5):159–168
- Pack S, Choi J, Kwon T, Choi Y (2007) Fast-handoff support in IEEE 802.11 wireless networks. *IEEE Commun Surv Tutor* 9:2–12

Handover

- [Handoffs, Modeling in IP Networks](#)

Handover in LTE-U/LAA

Jaewook Lee, Haneul Ko, and Sangheon Pack
Korea University, Seoul, South Korea

Synonyms

[Handoff in LTE-U/LAA](#); [Mobility management in LTE-U/LAA](#)

Definitions

- *Long-term evolution in unlicensed band (LTE-U):* LTE-U means that the 4G radio communications technology (i.e., LTE) is used in unlicensed spectrum such as the 5 GHz band, which is already used by 802.11 equipment.

Historical Background

To deal with the growing mobile data traffic, many researchers have paid attention to LTE-U (Huawei 2014; Qualcomm Technologies 2014). By extending an LTE network architecture to unlicensed spectrum and aggregating licensed spectrum and unlicensed spectrum by the carrier aggregation technique (Zhang et al. 2015), it is anticipated that LTE-U can efficiently mitigate the explosive mobile data traffic. However, the regulation in unlicensed spectrum should be complied to exploit unlicensed spectrum. The detailed regulations for using unlicensed spectrum are different according to countries. These regulations can be summarized as follows (3GPP 2015):

- *Transmission power:* LTE can control transmission power in unlicensed spectrum. In addition, the transmission power should be limited to the maximum value.
- *Location:* LTE can use unlicensed band for indoor only or both indoor and outdoor.

- *Dynamic frequency selection (DFS)*: LTE-U should apply the DFS mechanism designed to avoid causing interference to a non-international mobile telecommunication system (e.g., radar).
- *Listen before talk (LBT)*: LTE-U should apply the LBT mechanism for fair co-existence with other wireless communication systems (e.g., Wi-Fi).

In the United States of America and Korea, LBT implementation is not mandated to be used in unlicensed spectrum. Therefore, any device can access unlicensed spectrum without the LBT mechanism (i.e., clear channel assessment (CCA)) in those countries. On the contrary, in Europe and Japan, there is a mandatory requirement to implement LBT when accessing unlicensed spectrum. That is, when any device in those countries tries to access unlicensed spectrum, the device should conduct CCA beforehand. If the sensed energy level of unlicensed spectrum is lower than a certain threshold, the device can access the unlicensed band.

Due to the different regulations among countries, two types of LTE in unlicensed band have been proposed: (1) LTE-U and (2) Licensed Assisted Access (LAA). These are described as follows (4):

- *LTE-U*: LTE-U is proposed by Qualcomm and developed by the LTE-U forum. Many major companies such as Verizon, Ericsson, Qualcomm, and Samsung are included in this forum. Originally, LTE-U was designed for countries that do not have LBT regulation (e.g., the United States of America and Korea). LTE-U uses licensed spectrum as the primary carrier component for both control channel and data channel, whereas unlicensed spectrum is used only as the secondary carrier component to deliver data.
- *LAA*: LAA has been standardized by the 3GPP in Rel-13. The goal of LAA standardization is a unified global framework that complies with the regulation in the different regions of the world. Therefore, LAA is designed to satisfy the LBT requirement, which is mandated in some countries. Similar as LTE-U, LAA also uses licensed spectrum as the primary carrier component for signaling and delivering data, whereas the unlicensed spectrum is exploited as the secondary carrier component that is added to the primary carrier component by means of a carrier aggregation.

The detailed comparison between LTE-U and LAA is summarized in Table 1 (4).

Handover in LTE-U/LAA, Table 1 Comparison between LTE-U and LAA

Classes	LTE-U	LAA
Developed by	LTE-U forum	3GPP
Frequency bands	5150–5250 MHz and 5725–5850 MHz, whereas 5250–5725 MHz has been reserved for future use	5150–5925 MHz
Mode of operation	Supplemental downlink only	Downlink and up-link
Bandwidths allowed	20 MHz	10 and 20 MHz
Purpose	Use of the existing LTE specifications in countries that do not impose LBT	Single unified global framework that complies with regulations in the different world regions
Coexistence mechanisms	Carrier selection, on-off switching, carrier sensing adaptive transmission (CSAT)	Carrier selection and LBT
Main advantage	Will likely be the first version in the market	Will be a single unified global framework for LTE in unlicensed bands worldwide

Foundations

Since LTE-U/LAA aggregates licensed and unlicensed spectrum, there are several benefits compared to a Wi-Fi system, which can be summarized as follows (Zhang et al. 2015):

- *Significant boost in data rates:* Since LTE-U/LAA can use the carrier aggregation technique to aggregate licensed and unlicensed bands, higher throughput can be achieved. LTE-U/LAA is a synchronous system and adopts scheduling-based channel access, whereas Wi-Fi system applies contention-based random access. Therefore, higher spectrum efficiency (i.e., higher throughput) is guaranteed in an LTE system. Furthermore, multi-device frequency-selective diversity gain can be achieved based on the feedback of channel quality indicator (CQI) from devices.
- *Reliable and secure communication:* In LTE-U/LAA, the primary carrier component where control messages are transmitted is always set to licensed spectrum. This means that the control messages are exchanged properly between base station and devices regardless of unlicensed spectrum conditions. In addition, LTE-U/LAA has better authentication and authorization techniques than a Wi-Fi system, which provides more secure transmissions.
- *Seamless mobility and coverage:* In LTE-U/LAA, devices can access both licensed and unlicensed spectrum by the same LTE access procedures (e.g., integrated authentication and security). Moreover, the primary carrier component in licensed spectrum can always provide ubiquitous coverage. Therefore, when devices move out of the coverage of a base station, only the horizontal handover procedure (between same systems) is needed, and therefore the device's ongoing session can be switched as seamlessly as possible without any interruption.

Although LTE-U/LAA has quite a lot of advantages, LTE-U/LAA is still facing many challenges. One of the challenges is the mobility management. The handover procedure in the

original LTE networks is designed for the handover within licensed spectrum and can be explained as follows:

The handover procedure in the original LTE networks is divided into three phases (i.e., handover preparation phase, handover execution phase, and handover completion phase). In the original LTE networks, the UE periodically measures the radio signal strength of the neighbor cells and source cell. Then, the UE reports the strength information to its serving evolved node B (eNB) by using a Measurement Report message. If the signal strength of a certain neighbor cell is stronger than that of the serving cell, the serving eNB decides to execute the handover and initiates the handover preparation phase.

In the handover preparation phase, the source eNB sends a Handover Request message to the target eNB that has the strongest signal strength to the UE. When the target eNB receives the Handover Request message, the target eNB allocates a radio resource for the UE and replies a Handover Request ACK message to the source eNB. After receiving the Handover Request ACK message, the source eNB triggers the handover execution phase by sending a Radio Resource Control Connection Reconfiguration message to the UE.

In the handover execution phase, the UE disconnects from the source eNB and synchronizes to the target eNB. When the UE completes synchronization (e.g., random access process to associate the target eNB), the UE informs the target eNB by sending a Radio Resource Control Connection Reconfiguration Complete message, and the target eNB replies to the UE with an ACK.

Finally, in the handover completion phase, the path to the target eNB from the source eNB is switched in the LTE core network. When the path switch procedure is completed, the source eNB releases the UE context.

Since the handover procedure in the original LTE networks is designed for the handover within the licensed spectrum, if it is applied for the handover between licensed and unlicensed spectrum without any modification, the handover

performance can be degraded. When the UE is associated with the serving eNB and the conventional handover procedure is triggered to the target eNB with the unlicensed band, the delay of handover execution phase should be longer than that of the conventional handover with the licensed band. Since the other systems share the unlicensed band, the target eNB waits to access the unlicensed band until the other systems do not access the unlicensed band. Therefore, due to this additional delay, the overall delay of handover with the unlicensed band can be longer than that of the handover with the licensed band.

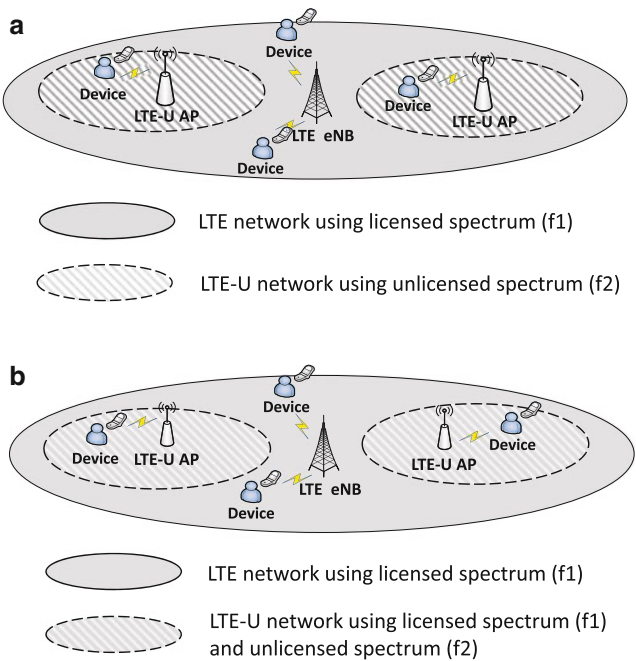
Lee et al. (2015) evaluated the handover performance of LTE-U/LAA in terms of TCP goodput by means of the NS-3 simulator with a modified LENA LTE module with the handover procedures in the original LTE networks. The evaluation result showed the degradation of the handover performance when the conventional handover procedure was applied without any modification.

In the simulation, specifically, the performance evaluation is conducted in two scenarios as shown in Fig. 1. In scenario 1, LTE evolved node B (eNB) and LTE-U access point (AP) transmit data through licensed spectrum (f1)

and unlicensed spectrum (f2), respectively. That is, the carrier aggregation technique is not supported. However, since unlicensed spectrum cannot be used as the primary carrier component for signaling, licensed spectrum (f1) from LTE eNB is used for signal channel even when a device is located in the communication region of LTE-U AP. On the other hand, in scenario 2, the LTE eNB can transmit data through licensed spectrum (f1), while the LTE-U AP can use both licensed spectrum (f1) and unlicensed spectrum (f2), and then it can aggregate them by means of carrier aggregation to transmit data. Since LTE eNB and LTE-U AP use the same licensed spectrum (f1), devices in the cell boundaries experience a co-channel interference. Meanwhile, the LTE-U AP should employ an LBT mechanism regardless of the scenarios. Also, regardless of the context, the UE measures the radio signal strengths of the neighbor cells and the source eNB and reports the radio signal strengths to the source eNB or AP. Then, if the signal strength of the neighbor cell is stronger than that of the serving cell, the source eNB or AP triggers the handover procedure. The following features are observed through extensive simulations.

Handover in LTE-U/LAA,

Fig. 1 Evaluation scenarios (a) Scenario 1. (b) Scenario 2



The handover performance (i.e., TCP goodput) of LTE-U/LAA can be degraded due to the additional access delay that is defined as the time right after the handover to data transmission and the co-channel interference. Note that the additional access delay is caused by the LBT mechanism. That is, since the LTE-U AP conducts CCA to occupy unlicensed spectrum (i.e., LBT mechanism), more delay is caused when devices handover to LTE-U AP. Moreover, when the LTE-U AP uses both licensed spectrum and unlicensed spectrum (i.e., scenario 2), the co-channel interference between LTE eNB and LTE-U AP occurs. Then, licensed spectrum quality is degraded, and thus the handover performance can be degraded. Meanwhile, compared to the conventional LTE where handover can only be triggered by the mobility of the device, the handover in LTE-U/LAA can also be triggered due to the unavailability of unlicensed spectrum in the serving cell. Therefore, devices should spend more time searching available unlicensed spectrum even though the received signal strength from the serving cell remains strong, which can degrade the handover performance. From the above results, we can conclude that the handover scheme for LTE-U/LAA should be devised with the following design objectives:

- *To reduce the additional access delay*
- *To mitigate the influence of co-channel interference*
- *To avoid unavailability-triggered handovers*

Key Application

In LTE-U/LAA, an appropriate handover procedure should be developed with the consideration of the characteristic of the unlicensed channel to avoid the performance degradation (e.g., the additional access delay and co-channel interference).

Cross-References

- [Handoffs, Modeling in IP Networks](#)
- [Mobility Management in 5G](#)

References

- 3GPP (2015) Study on licensed-assisted access to unlicensed spectrum; (Release 13). 3rd generation partnership project(3GPP), TS 36.889
- Huawei (2014) LTE for unlicensed spectrum. Huawei WhitePaper
- Labib M, Marojevic V, Reed JH, Zaghloul AI (2017) Extending LTE into the unlicensed spectrum: technical analysis of the proposed variants. arXiv preprint:1709.04458
- Lee J, Ko H, Pack S (2015) Performance evaluation of LTE- unlicensed in handover scenarios. In: Proceedings of the 2015 international conference on information and communication technology
- Nokia (2014) LTE-U: unlicensed spectrum utilization of LTE. Nokia White Paper
- Qualcomm Technologies (2014) LTE in unlicensed spectrum: harmonious unlicensed spectrum. Qualcomm White Paper
- Zhang R, Wang M, Cai LX, Zheng Z, Shen X, Xie LL (2015) LTE-unlicensed: the future of spectrum aggregation for cellular networks. IEEE Wirel Commun 22(3):150–159

Healthcare

- [Neighborhood-Based ICT-Driven Support \(NIS\) System of Emergency Medical Services \(EMS\) for Elderly People in Hyper-Aged Societies](#)

Healthcare Internet of Things

- [Internet of Medical Things](#)

Heterogeneous Cellular Networks

- [Resource Management in Macrocell–Small Cell Systems and D2D-Assisted Cellular Systems](#)

Heterogeneous Network Pricing

- [Oligopoly Pricing in Wireless Networks](#)

Heterogeneous Networks Through Multi-resources Deployment, Performance Enhancement for

Yuchen Zhou¹, Richard Fei Yu^{2,3}, Jian Chen¹, and Yonghong Kuo¹

¹School of Telecommunications Engineering, Xidian University, Shaanxi, China

²Carleton University, Ottawa, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

Resource management for heterogeneous networks

Definitions

The performance of heterogeneous networks can be greatly enhanced through a reasonable multi-resources deployment scheme.

Historical Background

Since the rapid evolution of terminals continually explode the traffic demand in wireless communication systems, the network providers are facing an enormous challenge to enhance the system capacity. As improvements in spectral efficiency (SE) at link level approach its theoretical limits with existing techniques, heterogeneous network (HetNet) architecture has been proposed as a promising solution to improve SE per unit area (Navaratnarajah et al. 2013). A HetNet may consist of different sizes of cells with different radio access technologies (RATs), such as Third Generation Partnership Project Long-Term Evolution (3GPP-LTE), Worldwide Interoperability for Microwave Access (WiMax), and Wireless Local Area Network (WLAN).

Generally, resource management plays a key role in implementing efficient multiuser com-

munications over complicated wireless network environments. Owing to its potential performance improvement, there have been extensive research efforts on the resource management strategy for wireless HetNet (Liu et al. 2017; Gerasimenko et al. 2017). Apart from the capacity demand, considering energy efficiency (EE) is also of paramount importance for sustainable growth due to the increasing network usage of the latest advanced wireless communication devices (Chen et al. 2016). This section will indicate several possible research directions for the resource management in HetNets.

Business Diversity

Introduction

Video content traffic in the next-generation mobile cellular networks is estimated to exponentially grow, owing to the surge in requirements of bandwidth-intensive applications based on video streaming. To support the massive content delivery, caching and multicast are valuable to be exploited for the next-generation wireless networks (Poularakis et al. 2016). Note that recent advantages of *information-centric networking* (ICN) (Fang et al. 2015) instead of simply caching at base stations (BSs) can further facilitate the dissemination of information, since ICN is characterized by in-network caching that enables the network nodes to not only store but also name the contents passing through them, thus promoting the findability and delivery of the information. However, it is still worth noting that a video content is possibly demanded to be transcoded into several versions for adapting the complex network conditions and meeting the various device requirements. To promote computing capability of BSs, *mobile edge computing* (MEC) (Kumar et al. 2016) as a promising technique is necessary to be integrated at BSs for realizing video transcoding, since it is able to support agile and pervasive computation services through enabling cloud computing capability at BSs. Furthermore, a joint optimization of transcoding, caching, and multicast is crucial, since the process value and

the result value of both caching and computing depend on communications.

Wireless network virtualization has also attracted great attentions, since it is able to significantly reduce the capital expenses and the operational expenses of wireless access networks as well as core networks (Liang and Yu 2015). Through virtualization, wireless network infrastructure can be decoupled from their provided services, and various users with differentiated services requirements can dynamically share the same infrastructure, thereby maximizing the system utilization.

Wireless network virtualization provides a better platform to install ICN and MEC functions in HetNets. With an effective resource management strategy, the diversity demands of user services can be better satisfied over the virtualized HetNets networks based on ICN and MEC. Therefore, this subsection focuses on the combination of ICN and MEC in virtualization and the efficient allocation of the limited resources including the computing, caching, and communication resources.

A Feasible Multi-resources Deployment Plan

Figure 1 illustrates the proposed HetNets virtualization framework for transcoding, caching, and multicast. The virtualization procedure can convert such a physical HetNets with abundant physical resources to several virtual networks. Compared to the traditional networks, where the whole system is designed to support different kinds of services, the virtualized HetNets can focus on only one type of service and bring in a better user experience. As shown in Fig. 1, two virtual networks are created according to the quality-of-service (QoS) requirements of multicast groups and the properties of contents. The personal services, such as text, and the real-time services, such as voice call and video meeting, can be executed in virtual traditional HetNets without caching and computing functions, i.e., virtual network 2, since the contents of this kind of services are private and of low reuse probability. The virtual ICN-/MEC-enabled HetNets, i.e., virtual network 1, are more specified on video streaming services. Assume that the video

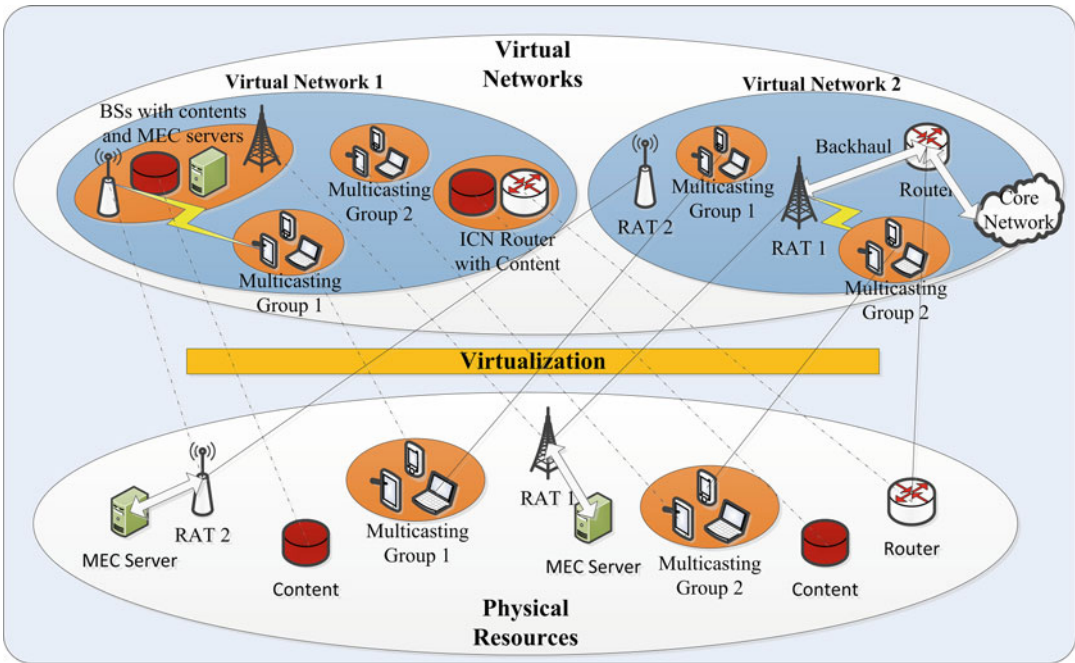
contents required by different multicast groups are encoded into several versions with different bitrates to adapt the complex network conditions and to meet the various device requirements. With the help of ICN, the video content versions with higher popularity can be cached at BSs for alleviating the backhaul usage. With the help of MEC, some of the video versions are not necessary to be cached and can be directly transcoded from the already cached versions.

In this way, there are two issues to be solved at each BS: (i) For a content, which versions are valuable to be stored? (ii) Which versions are chosen to be computed (transcoded)? The solutions of these two questions result in two kinds of alternative costs: caching cost and computing cost. Note that these two kinds of costs are competing with each other. If a BS has a larger storage capacity, the more video versions can be cached, so that the probability of computing will be reduced, the result of which leads to a lower computing cost. In contrast, if a BS is assigned more computing resource, the less video versions need to be cached, so that the storage space can be saved, the result of which leads to a lower caching cost. In conclusion, the resource management strategy should have the ability to jointly optimize the utilities of caching, computing, and communications, through selecting the appropriate virtual network for each mobile user and storing the proper versions at each BS.

Communication Reliability

Introduction

WLAN/femtocell interworking system, as a typical example of HetNets, has been recognized as a promising technique to cater for the ever-increasing service demand (Gamage et al. 2014). Since the complicated network environment is always with the high operation overhead, self-organizing network (SON) has been proposed to realize the automated network management and the low operation cost (Barco et al. 2012). As one of the main functions of SON, self-healing is capable of detecting, diagnosing, and compensating the abrupt network failure.



Heterogeneous Networks Through Multi-resources Deployment, Performance Enhancement for, Fig. 1
Illustration of HetNets virtualization framework with ICN and MEC

Note that in a real scenario, the same region is usually covered by a number of networks with different sizes and RATs. While implementing the self-healing functions for such a real scenario, not only the parameters but also the methods have to be designed to apply to each specific network (Zhou et al. 2017). Therefore, it is quite tricky but worthy of designing a resource management strategy for self-healing, which can be well applied in such a complicated scenario.

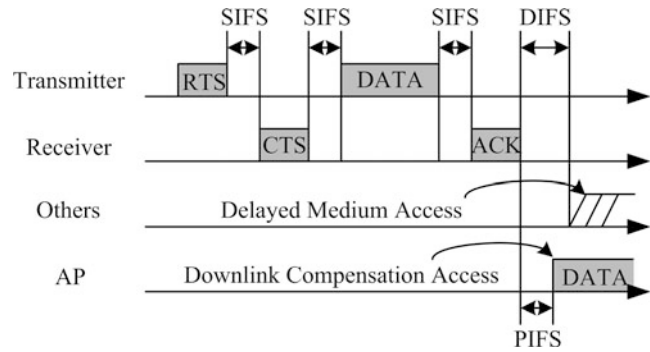
A Reliable Resources Deployment Plan

The key challenge for self-healing in WLAN/femtocell system is the high complexity caused by existence of multiple medium access control layer (MAC) and physical layer (PHY) techniques. It should be noticed that the feasible transmission rate of the interworking system depends on MAC and PHY techniques of the different networks (Gamage et al. 2014). Therefore, to enable self-healing for the interworking system, the proposed resource management scheme is designed based on MAC of IEEE

802.11 WLANs and PHY of femtocells. WLANs and femtocells are connected to a centralized unit viewed as operation and management (OAM). Assume that faulty networks have already been detected by any outage detection method, and the users in faulty networks, named as compensation users, can be continually served by neighbor normal WLANs or femtocells with the help of OAM.

WLANs usually adopt the distributed coordination function (DCF) mechanism with request-to-send (RTS)/clear-to-send (CTS) handshaking. Figure 2 illustrates the DCF mechanism with the downlink compensation access. In the traditional DCF-based WLAN system, carrier sense multiple access with collision avoidance (CSMA/CA) protocol can effectively avoid the collision of data frames. However, this kind of protocol cannot guarantee the required utilization of the downlink transmission, especially when a large number of users access the WLAN. Thus, a required utilization ratio ψ is introduced and defined as the ratio of the downlink throughput to the uplink

Heterogeneous Networks Through Multi-resources Deployment, Performance Enhancement for, Fig. 2
DCF mechanism with downlink compensation access



throughput, to ensure an adequate level of the fairness between the downlink and the uplink. If the current utilization ratio is less than ψ , AP can preferentially transmit data frames without collision after point interframe space (PIFS), since PIFS is shorter than DCF interframe space (DIFS). In this way, while implementing self-healing for WLANs, the downlink utilization can be well guaranteed, for the users within the normal networks and fault networks.

For the femtocells, a joint subcarrier and power allocation scheme is usually adopted for performance optimization. However, many existing works are based on the assumption that channel gain from transmitter to receiver is perfectly estimated. In practice, it is really difficult and expensive to accurately estimate instantaneous channel between transmitter and receiver, because of channel fluctuations, limited feedback capacity, and channel estimation and quantization errors (Fang et al. 2017; Chang et al. 2017). Inaccurate channel estimation results in imperfect channel state information (CSI), which leads to the collision of users and the outage of networks. Therefore, the resources need to be managed for satisfying the acceptable outage probabilities and restricting the collision probabilities for each user.

The optimization problem for self-healing can be designed to jointly optimize the allocations of users, subcarriers, and power, which is a nonconvex mixed-integer nonlinear problem (MINLP). To solve the MINLP suboptimally but effectively, the compensation users can be allocated

to normal WLANs firstly, since the co-channel interference issue needs to be addressed in the resource optimization of femtocells. If the users in WLANs tend to be saturated, the remaining compensation users will be allocated to normal femtocells, and then the joint subcarrier and power allocation scheme can be executed for performance optimization.

Key Applications

A reasonable multi-resources deployment strategy is able to promote the network performance. In addition, the business diversity and the communication reliability are two major issues that could be solved through reasonable multi-resources deployment. Therefore, the resources deployment plays a key role in implementing efficient multiuser interactions over different wireless networks, such as cellular network, femtocell, WLAN, and other environments.

Cross-References

- [Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks](#)
- [Integrated System of Networking, Caching, and Computing](#)

► **Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks**

References

- Barco R, Lazaro P, Munoz P (2012) A unified framework for self-healing in wireless networks. *IEEE Commun Mag* 50(12):134–142
- Chang Z, Wang Z, Guo X, Han Z, Ristaniemi T (2017) Energy-efficient resource allocation for wireless powered massive mimo system with imperfect CSI. *IEEE Trans Green Commun Netw* 1(2):121–130
- Chen J, Zhou Y, Kuo Y (2016) Energy-efficiency resource allocation for cognitive heterogeneous networks with imperfect channel state information. *IET Commun* 10(11):1312–1319
- Fang C, Yu FR, Huang T, Liu J, Liu Y (2015) A survey of green information-centric networking: research issues and challenges. *IEEE Commun Surv Tutorials* 17(3):1455–1472
- Fang F, Zhang H, Cheng J, Roy S, Leung VC (2017) Joint user scheduling and power allocation optimization for energy-efficient noma systems with imperfect CSI. *IEEE J Sel Areas Commun* 35(12):2874–2885
- Game AT, Liang H, Shen X (2014) Two time-scale cross-layer scheduling for cellular/wlan interworking. *IEEE Trans Commun* 62(8):2773–2789
- Gerasimenko M, Moltchanov D, Andreev S, Koucheryavy Y, Himayat N, Yeh S, Talwar S (2017) Adaptive resource management strategy in practical multi-radio heterogeneous networks. *IEEE Access* 5:219–235
- Kumar N, Zeadally S, Rodrigues JJ (2016) Vehicular delay-tolerant networks for smart grid data management using mobile edge computing. *IEEE Commun Mag* 54(10):60–66
- Liang C, Yu FR (2015) Wireless network virtualization: a survey, some research issues and challenges. *IEEE Commun Surv Tutorials* 17(1):358–380
- Liu C, Li M, Hanly S, Whiting P (2017) Joint downlink user association and interference management in two-tier HetNets with dynamic resource partitioning. *IEEE Trans Veh Technol* 66(2):1365–1378
- Navaratnarajah S, Saeed A, Dianati M, Imran MA (2013) Energy efficiency in heterogeneous wireless access networks. *IEEE Wirel Commun* 20(5):37–43
- Poularakis K, Iosifidis G, Sourlas V, Tassiulas L (2016) Exploiting caching and multicast for 5g wireless networks. *IEEE Trans Wirel Commun* 15(4):2995–3007
- Zhou Y, Chen J, Kuo Y (2017) Cooperative cross-layer resource allocation for self-healing in interworking of wlan and femtocell systems. *IEEE Commun Lett* 21(1):136–139

Heterogeneous Small Cell Networks

Haijun Zhang¹, Xiaoli Chu², Weisi Guo³, Siyi Wang^{4,5}, and Keping Long¹

¹Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services, University of Science and Technology Beijing, Beijing, China

²Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, UK

³School of Engineering, University of Warwick, Coventry, UK

⁴Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China

⁵The Institute for Telecommunications Research, University of South Australia, Adelaide, Australia

Introduction

In recent years, the use of mobile data has grown by 70% to 200% annually. What is more worrying is that the bursty nature of wireless data traffic makes traditional network capacity planning obsolete (Zhang et al. 2018). Almost all operators and suppliers believe that small cells (such as picocells, femtocells, and relay nodes) are a promising solution to improve local capacity in traffic hotspots, thereby relieving macrocell burden. In order to effectively divert macrocell traffic to small cells, researchers have done a lot of research, mainly in indoor situations (Hwang et al. 2013).

Licensed spectrum is scarce in cellular networks. However, deploying small cells within the coverage of macrocells also hopes to obtain the same spectrum as macrocells (Chu et al. 2013). A frequency reuse factor of 1 in 3G HSPA+ and 4G LTE/LTE-A systems has proven to yield high gains in network capacity. If there is no rich extra spectrum for mobile communications, future cel-

lular networks will have to actively explore new and more efficient methods of frequency reuse. Therefore, it is highly probable that the ultra-intensively deployed small cell may be affected by interference between the small cell and macro-cell co-channel and adjacent small cells.

According to the current situation, both industry and academia are actively studying all possible spectrum resources, including unlicensed frequency bands, to provide more timely and quality services (Huawei 2013). Industry believes that the unlicensed 2.4 GHz and 5 GHz bands in Wi-Fi systems are important candidates for providing additional frequency bands for mobile communications. An unlicensed band originally targeted at 5 GHz may have up to 500 MHz of spectrum available.

At present, the vast majority of mobile devices, such as mobile phones and tablets, support Wi-Fi access, and more Wi-Fi access points will be available in the future (Zhang et al. 2015). Especially in developed cities, the density of Wi-Fi access points has reached more than 1,000 per square kilometer. Such a widely deployed Wi-Fi system plays a decisive role in the data traffic diversion of the cellular network and is more prominent in indoor traffic hotspots and poor signal areas. As the FCC decided to use 100 MHz of spectrum in the 5 GHz band for Wi-Fi use, operators have more opportunities to expand data services to Wi-Fi systems. Wi-Fi access points have even been regarded as a distinct tier of small cells in heterogeneous cellular networks. At present, the industry is exploring the use of unlicensed spectrum of LTE/LTE-A under the Wi-Fi system. The coexistence and interworking of Wi-Fi and heterogeneous cellular networks will become a research hotspot and research trend in the future. If the interference between Wi-Fi and cellular networks can be effectively mitigated, the joint deployment of Wi-Fi and cellular networks under unlicensed spectrum will significantly increase the overall system capacity of heterogeneous networks.

Because of the huge benefits of joint deployment of Wi-Fi and cellular networks under unlicensed spectrum, more and more research groups are starting to join. In Hajmohammad

and Elbiaze (2013), the authors proposed a quality of service (QoS)-based strategy to split the unlicensed spectrum between Wi-Fi and femtocell networks. Although the unlicensed spectrum splitting scheme considers fairness between Wi-Fi access points and femtocells, the split use of the spectrum between two systems prohibits a high crossnetwork throughput. In Charitos and Kalivas (2013), the authors discussed the deployment of a heterogeneous vehicular wireless network consisting of IEEE 802.11b/g/e Wi-Fi and IEEE 802.16e WiMAX systems inside a tunnel for surveillance applications and specifically evaluated the handover performance of the hybrid Wi-Fi/WiMAX vehicular network in an emergency situation. In Almeida et al. (2013), time-domain resource partitioning based on the use of almost blank subframes (ABSs) was proposed for LTE networks to share the unlicensed spectrum with Wi-Fi systems. Qualcomm has recently proposed to deploy LTE-A in the unlicensed 5 GHz band currently used mostly by Wi-Fi. The main idea is to deploy LTE-A as supplemental downlink (SDL) in the 5725–5850 MHz band in the USA, with the primary cell always operating in the licensed band. Verizon and Ericsson are also exploring similar ideas. Huawei and CMCC have investigated the availability, commonality, and feasibility of integrating the unlicensed spectrum to International Mobile Telecommunications-Advanced (IMT-A) cellular networks (Huawei 2013). On pre-commercial multicell networks, LAA-LTE achieves better coverage and capacity in the 5 GHz unlicensed spectrum than Wi-Fi alone have been proved. So there are worries that LTE-U will occupy all the bandwidth of Wi-Fi in ultradense deployment.

There are still many technical problems with the coexistence of Wi-Fi, and cellular networks under unlicensed spectrum, such as efficient spectrum sharing (Zhang et al. 2014) and interference mitigation (Zhang et al. 2016), have not been sufficiently addressed. In Dimatteo et al. (2011), the authors proposed an integrated architecture exploiting the opportunistic networking paradigm to migrate data traffic from cellular networks to metropolitan Wi-Fi access

points. Bennis et al. (2013) introduced the basic building blocks of cross-system learning and provided preliminary performance evaluation in an LTE simulator overlaid with Wi-Fi hotspots. For the unlicensed spectrum sharing deployment of Wi-Fi and LTE-A systems, the co-channel interference between Wi-Fi tier and LTE-A tier can be mitigated by using ABSs, in which the interfering tier is not allowed to transmit data, and the victim tier can thus get a chance to schedule transmissions in the ABSs with reduced cross-tier interference (Liang et al. 2012). Moreover, it has been demonstrated that by estimating the number of co-channel transmitters and knowing the deployment density of network nodes in a region, the average channel quality at any point in a coverage area can be inferred (Wang et al. 2014).

In this entry, we propose a system architecture that supports coexistence of Wi-Fi and heterogeneous cellular networks to share unlicensed spectrum. Based on the system architecture, we first provide an in-depth review of the ABS scheme used for mitigating the co-channel interference from small cells to Wi-Fi systems and propose a spectrum-sensing-based fair ABS scheme without priority. We then present an interference avoidance scheme based on small cells estimating the density of nearby Wi-Fi transmissions to facilitate the unlicensed spectrum sharing between small cells and Wi-Fi access points. Simulation results are provided to evaluate the performance of the proposed system architecture and interference avoidance scheme in facilitating coexistence of Wi-Fi and 4G heterogeneous cellular networks in unlicensed spectrum.

Network Architecture for Wi-Fi and Cellular Coexistence

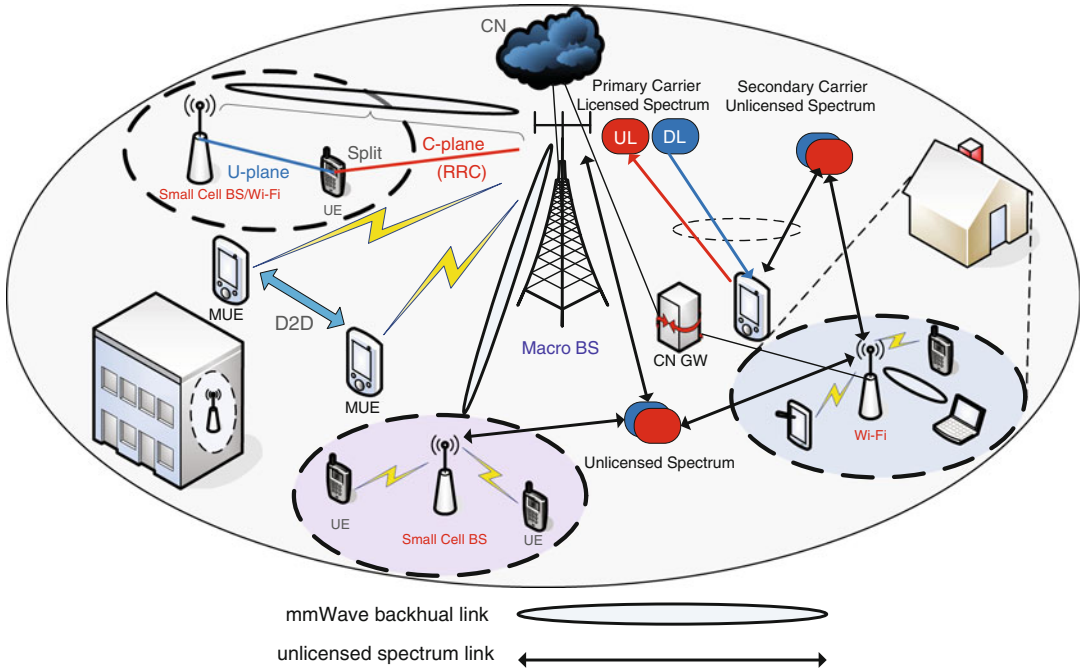
In this part, we research the system architecture where several Wi-Fi access points and small cells coexist in the coverage area of a macrocell.

Heterogeneous Network Architecture

As shown in Fig. 1, multiple Wi-Fi access points and small cells coexist in a macrocell coverage area in the system structure. The macrocell, small

cells, and Wi-Fi access points share the same unlicensed spectrum for providing radio access to users, millimeter-wave radio is used for small cell backhaul links, and device-to-device (D2D) communications are supported based on Wi-Fi Direct or LTE Direct. As shown in Fig. 1, the control plane (C-plane) and user plane (U-plane) are split on the radio links associated with small cells. Specifically, the C-plane of user equipments (UEs) associated with a small cell is provided by the macro eNB in a low-frequency band, while the U-plane of UEs associated with a small cell is provided by their serving small cell in a high-frequency band. For UEs associated with the macrocell, both the C-plane and U-plane of their radio links are provided by the serving macrocell. Since the C-plane of small cell UEs is managed by the macrocell, the radio resource control (RRC) signalings of small cell UEs are transmitted from the macrocell, and the handover signaling overhead between small cells and the macrocell can be much reduced (Nakamura et al. 2013). Regarding Wi-Fi access points as a type of small cells, the C-plane and U-plane split can be applied in a Wi-Fi/macrocell scenario, where the C-plane of UEs associated with a Wi-Fi access point is provided by the macro eNB in a low-licensed-frequency band, while the U-plane of UEs associated with a Wi-Fi access point is provided by their serving Wi-Fi access point in a high-unlicensed-frequency band. The interworking of Wi-Fi and cellular networks benefits from the split of C-plane and U-plane in terms of mobility robustness, service continuity, reduction in cell-planning efforts, energy efficiency, etc.

In an ultra-intensively deployed small cell, expensive wired backhaul is not feasible when considering the small cell backhaul problem. In the meanwhile, in-band wireless backhaul solutions using the licensed spectrum may not be feasible either, because of the scarcity of the licensed spectrum (Sooyoung et al. 2013). Recently, the millimeter-wave bands, such as the unlicensed 60 GHz band and the low interference licensed 70 GHz and 80 GHz bands, have been considered as promising candidates for the small cell backhaul solution (Zhang et al. 2017a). This is motivated by the huge frequency bandwidth (globally har-



Heterogeneous Small Cell Networks, Fig. 1 Network architecture of LTE-A and Wi-Fi coexistence

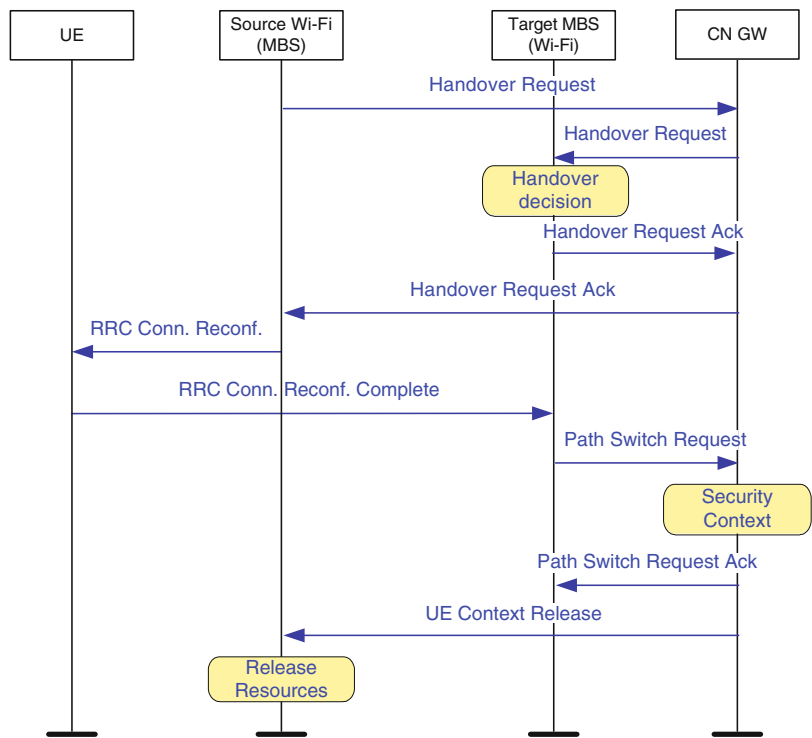
monized over more than 6 GHz millimeter-wave spectrum) that can be exploited and the spatial isolation supported by highly directional beams. At the same time, the new IEEE 802.11ad standard, a.k.a. WiGig, uses the unlicensed 60 GHz millimeter-wave band to deliver data rates of up to 7 Gbps. This adds a large amount of new frequency bandwidth to existing Wi-Fi products, such as 802.11n operating in the 2.4 GHz and 5 GHz bands and 802.11ac operating in the 5 GHz band.

For providing radio access, the macrocell, small cells, and Wi-Fi access points partake the unlicensed spectrum. The network architecture in Fig. 1 integrates the coexistence of Wi-Fi and cellular networks and facilitates smart management of data traffic in mobile operators' networks. For instance, the data traffic could be dynamically routed to the optimal radio interface for a particular application and user, with network congestion, reliability, security, and connectivity cost taken into account. As can be seen from Fig. 1, the primary carrier always uses licensed spectrum to transmit control signaling, user data, and mobility signaling, while

the secondary carrier(s) use unlicensed spectrum to transmit best-effort user data in the downlink and potentially the uplink.

Handover Procedure Between Wi-Fi and LTE/LTE-A

One of the key aspects of interworking between Wi-Fi and LTE/LTE-A systems is seamless mobility. In order to ensure the user's QoS during the handover, there should be no packet loss or radio link failure. Current Wi-Fi systems do not support interworking between Wi-Fi and LTE/LTE-A systems, but this is necessary because users often move between Wi-Fi access points and cellular network coverage areas. In 3GPP Rel-11, trusted WLAN gain, the Enhanced Packet Core (EPC) is built on S2a-based Mobility over GPRS Tunnelling Protocol (SaMOG), which is enhanced in 3GPP Rel-12 to provide traffic steering and mobility between LTE-A and Wi-Fi networks and to optimize the use of network resources. The seamless and non-seamless handover solution between 3GPP and non-3GPP access networks is the 3GPP standard



Heterogeneous Small Cell Networks, Fig. 2 Handover procedure between Wi-Fi and LTE/LTE-A

TS 23.401. The role of these standards is to enable users to continue using data services while passing through macrocells, small cells, and Wi-Fi hotspots. In the network architecture shown in Fig. 1, a UE can switch between Wi-Fi and cellular networks through the core network (CN) gateway (GW). In Fig. 2, the CN GW-based handover procedure between a source Wi-Fi (macrocell) and a target macrocell (Wi-Fi) is given. Since the C-plane of the Wi-Fi user equipment is managed by the macrocell, the RRC signaling of the Wi-Fi user equipment is sent by the macrocell, so that the signaling overhead between the Wi-Fi access point and the macrocell can be greatly reduced (Zhang et al. 2015).

Almost Blank Subframe Allocation

This section discusses almost blank subframe (ABS) schemes and does not have priority to mitigate the co-channel interference of small cells with Wi-Fi systems.

Coexistence Without Priority

In recent times, regulators are considering the possibility of coexisting multiple different radio access technologies (RATs) in the same frequency band. These include licensed bands (e.g., television spectrum) and unlicensed bands (e.g., amateur spectrum). At present, the USA, Canada, and the UK are the main countries, and supervision is under way to allow the operation of space equipment (WSDs), for example, the IEEE 802.22 fixed point-to-point cognitive radio transmissions in television white space and more lately the IEEE 802.16h wireless broadband protocols.

The secondary user uses spectrum sensing technology in the licensed band to prevent interference with the primary user because the primary user and the secondary user have a clear concept. This can be obtained by identifying the primary transmission using spectrum sensing or geo-location database operations. However, there is currently a lack of research activities to investigate how auxiliary users associated with different

RATs prevent or mitigate co-channel interference between each other. In an unlicensed band with no primary user and no secondary user concept, there are similar challenges between different RATs sharing the same spectrum.

In this chapter, we relate to the coexistence of two RATs without priority ranking as *coexistence without priority*. More specifically, in order to solve the problem of inconsistency between the primary user and the secondary user, we have focused on allowing cellular communications to coexist with Wi-Fi communications on an equal footing. The new content here refers to the coexistence of two different RATs and the impact on the interference graph when these two RATs do not coexist. Although the coexistence of contention-based systems has been explored (e.g., 802.11 and 802.15 systems) (Yuan et al. 2010), the absence of prioritization between non-competitive systems (LTE) and contention systems (Wi-Fi) is a bad exploration for competing systems. In fact, we can think that it is entirely reasonable to suspect that LTE's allocation-based transport protocol may completely prevent the conflict-based Wi-Fi protocol. In addition, the density of small cells is getting higher and higher, and this interpretable defect in the same frequency band may cause serious capacity problems.

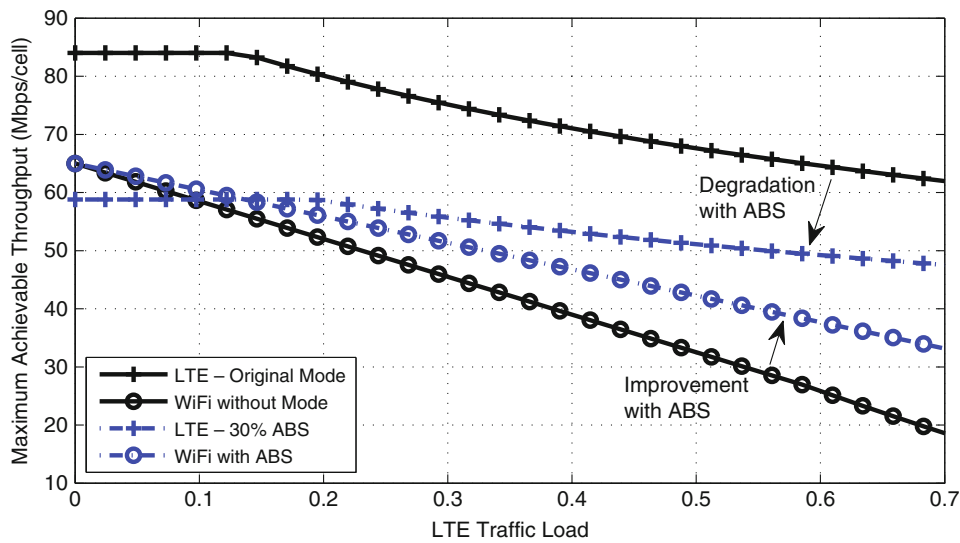
The communication protocol in the coexistence of multiple RATs can operate in its default normal mode or coexistence mode. The latter is triggered when proximity detects another RAT and requires action. Coexistence mechanisms can be divided into two groups: (i) those that require message exchanges between nodes or RATs and (ii) those that are not needed. In general, cross-RAT coordination is very difficult because of different protocol development processes and vendor differences. Therefore, in the next section, we will review the non-collaborative coexistence mechanism that allows LTE to coexist with Wi-Fi in the unlicensed spectrum.

Random Almost Blank Subframe Allocation

In Almeida et al. (2013), autonomous (uncoordinated) coexistence between LTE and Wi-Fi

was achieved through LTE transmission of ABSs under a 3GPP Rel-10 time division duplex (TDD) scheme. ABS is a subframe of power or content reduction. Because several synchronization channels (e.g., common reference signals) are reserved, they are backward compatible with 3GPP Rel-8 and Rel-9. The ABS is triggered by the coordination message between the eNodeBs through the X2 interface to avoid interference between the cells of the same RAT. The frequency of ABS transmission can adapt to the time-varying interference environment. For the foregoing reasons, in order to avoid interference between cells of different RATs, coordinated messaging between cells of different RATs is a challenging task. Therefore, ABS is sent randomly at a certain speed, does not require coordination, and does not require backward compatibility with previous releases of undistributed bands (Almeida et al. 2013). The core idea of this idea is that in a randomly acquired ABS, the Wi-Fi access point can detect that the channel is idle and transmit after its contention-based protocol. Based on this, the allocation properties of LTE transmissions can be suppressed in a way that avoids coordination or spectrum sensing, but this comes at the price of reduced LTE spectrum efficiency and network capacity.

In Fig. 3, the maximum achievable throughput for an LTE or Wi-Fi (802.11n) cell under different normalized traffic loads is plotted, with LTE operating either under the original mode or with 30% of the subframes randomly selected as ABSs. The parameters used in the simulation can be found in Table 1. The results show that in the original mode, the LTE cell capacity is saturated at about 84 Mbps and the Wi-Fi cell capacity is saturated at 64 Mbps. The reason for this difference is due to the use of a discrete modulation and coding scheme (64QAM turbo code). As the LTE traffic load increases, the capacity of all cells decreases due to increased radio resource usage and the resulting increased interference. The LTE cell capacity attenuation is slower than the Wi-Fi capacity. The *superlinear* degradation of Wi-Fi capacity may finally lead to 0 Mbps. By employing 30% ABSs in LTE, the Wi-Fi capacity is improved significantly at high



Heterogeneous Small Cell Networks, Fig. 3 Maximum achievable throughput (Mbps per cell) for an LTE or Wi-Fi cell under different normalized traffic loads

(normalized to cell capacity), with LTE operating with (a) original mode and (b) ABS

Heterogeneous Small Cell Networks, Table 1 Simulation Parameters

Parameter	Value
LTE carrier frequency	2100 MHz
Wi-Fi carrier frequency	2400 MHz
LTE cell density	3 per km ²
Wi-Fi density	300 per km ²
LTE cell transmit power	40 W
Wi-Fi transmit power	1 W
Pathloss	3GPP urban micro
Peak LTE throughput	84 Mbits/s (64 QAM SISO)
Peak 802.11n throughput	65 Mbits/s (64 QAM SISO)

ness among networks, but ABS mechanism will decrease the capacity of LTE and Wi-Fi networks.

Interference Avoidance with Neighborhood Wi-Fi Density Estimation

This part researches an interference avoidance scheme based on small cells estimating the density of nearby Wi-Fi access points to facilitate their coexistence while sharing the same unlicensed spectrum.

LTE traffic loads, while the LTE cell capacity falls by 10–24 Mbps. It should be noted here that, with LTE ABS transmission, the capacity reduction of LTE and Wi-Fi will be slowed down because the random ABS can reduce cross-layer and interlayer intercell interference. It is worth noting that the overall total capacity of LTE and Wi-Fi is actually reduced by ABS, which indicates that the random ABS mechanism is more conducive to fairness than to the overall capacity.

In conclusion, LTE transmitters can randomly transmit ABSs when LTE and Wi-Fi share same spectrum and ABS mechanism can benefit fair-

Inference Framework

It can be seen that avoiding co-channel interference in a network with a high interference intensity can enhance the long-term system capacity (Liang et al. 2012; Zhang et al. 2017b). However, coordinating interference avoidance on the radio resource management (RRM) level needs a large volume of coordination information from multiple base stations (BSs) through the X2 interface. Especially, every BS requires to know whether its adjacent BSs are transmitting on each available radio resource block. This level of coordination taxes the *backhaul capacity*, while any *delay* in

information sharing may cause the interference avoidance performance to falter.

In Wang et al. (2014), an interference estimation technique that does not require information sharing between BSs and UEs was designed based on BSs sensing the spectrum and estimating the number of co-channel transmissions in a defined observation region. By estimating the number of co-channel transmitters and receiving the information about cell density in the district, the average channel quality at *any random point* in a coverage area can be inferred. As the expressions are tractable, the computational complexity is very low. This methodology can be applied to a K-tier heterogeneous network by leveraging a stochastic geometry framework and an opportunistic interference reduction scheme, which was shown to approach the interference estimation accuracy achievable by information exchange on the X2 interface (Wang et al. 2014).

The inference framework is based on the assumption that all cells have spectrum sensing equipment (low-cost spectrum sensing equipment for 2–5 GHz is now readily available). On each frequency band f , each sensor in cell (located at distance h from the BS) is capable to detect the power density P_f from all co-channel transmitters in an unbounded zone. Based on the spatial distribution of co-channel cell deployments (Haenggi 2005), the density of co-channel transmissions λ_f can be inferred from the P_f measurements (Wang et al. 2014):

$$\lambda_f \propto \frac{\sqrt{P_f/P}}{Q(h, \alpha)} \quad (1)$$

where α is the pathloss distance exponent, P is the average transmit power of the BSs, and the function $Q(h, \alpha)$ is given in Wang et al. (2014). This inference framework can be applied to a K-tier heterogeneous network which consists of macrocells, femtocells, and Wi-Fi access points. Figure 4 shows a femtocell sensing the received power spectrum to infer the number of co-channel transmitters in a three-tier heterogeneous network. For example, the estimated transmitter activities on the considered frequency band are 50% of LTE macrocells, 100% of LTE

femtocells, and 25% of Wi-Fi access points. This spectrum sensing mechanism is not capable to know which cells are transmitting, but it provides a statistical notion for a BS to infer the channel quality of a served user.

Based on the inferred density of co-channel transmissions λ_f in the vicinage, the signal-to-interference ratio (SIR) on frequency band f at distance d away from the sensing BS is estimated as:

$$\text{SIR}_{f,d} \propto P_f^{-1} \left[\frac{Q(h, \alpha)}{Q(d, \alpha)} \right]^2, \quad (2)$$

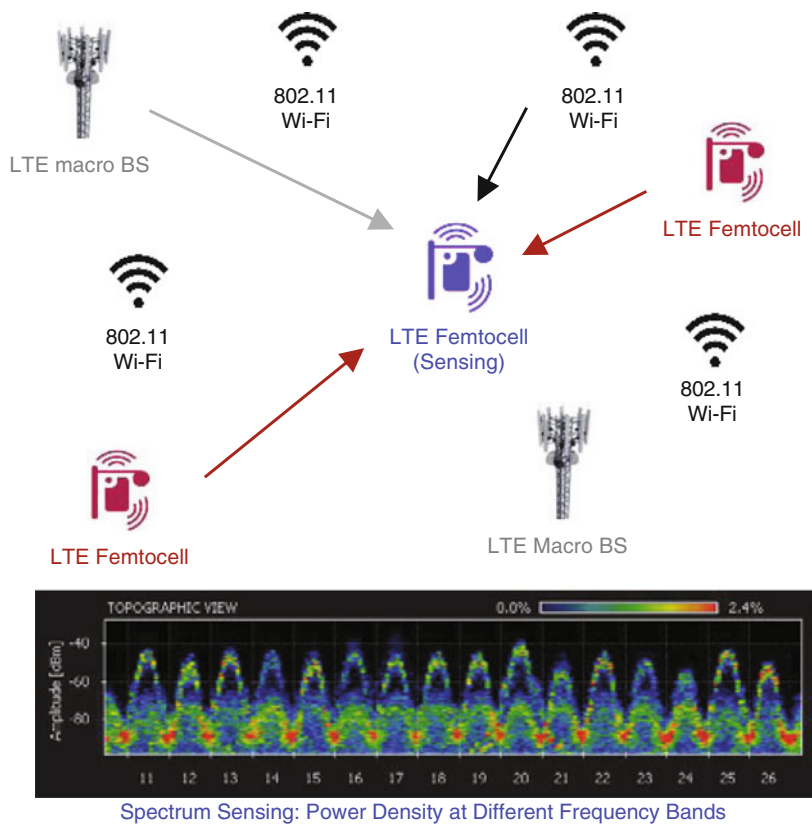
where the constant of proportionality is the received signal strength.

Simulation Results

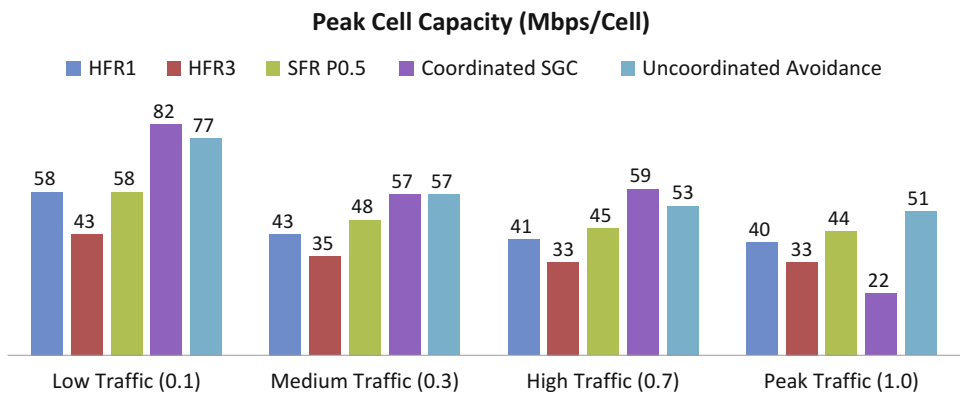
Figure 5 shows the simulation results of peak cell throughput versus normalized cell traffic load for a variety of static and dynamic interference mitigation schemes (Wang et al. 2014; Turyagyenda et al. 2012). The baseline is a hard frequency reuse 1 (HFR1) scheme, which shows a peak velocity of 58 Mbps/cell when the cells are unloaded (i.e., minimum intercell interference). This value will fall to 40 Mbps/cell for fully loaded cells without interference mitigation (i.e., maximum intercell interference). A similar trend exists for HFR3. The soft frequency reuse with a power back-off factor of 0.5 (SFR P0.5) performs better than the previous two schemes. We can see from Fig. 5 that the TDD-based sequential game coordinated (SGC) interference avoidance scheme achieves a much higher peak cell capacity than the HFR and SFR schemes at low and medium cell traffic loads. The uncoordinated interference avoidance scheme proposed in Wang et al. (2014) provides the highest peak cell velocity at high traffic loads and achieves over 90% the peak cell capacity of the SGC interference avoidance scheme which requires channel state information.

Discussion and Challenges

The disadvantage with a TDD-based coordinated interference avoidance scheme is the need for two prerequisites (Turyagyenda et al. 2012): (1) cell



Heterogeneous Small Cell Networks, Fig. 4 Illustration of a femtocell inferring the number of co-channel transmitters from a three-tier cellular and Wi-Fi heterogeneous network by sensing the received power spectrum (Wang et al. 2014)



Heterogeneous Small Cell Networks, Fig. 5 Plot of peak cell capacity versus different normalized cell traffic loads for a variety of static and dynamic interference mitigation schemes (Wang et al. 2014; Turyagyenda et al. 2012)

pairing or clustering and (2) static or dynamic assignment of cell priority. Effective cell pairing often involves the association of cells that are dominant interferers to each other. However, this may not always be the case. For example, the antenna boresight of BS A is pointing at BS B,

but the antenna boresight of BS B is pointing at a direction away from BS A . Alternatively, two cells that are closest to each other will be paired together. Cell priority assignment refers to the process of assigning different transmission priorities to cells. Random access, traffic-weighted, and QoS-weighted cell priority assignments have been considered in the literature.

Conclusion

In this entry, we have investigated the potentials and challenges associated with coexisting Wi-Fi systems and heterogeneous cellular networks sharing the unlicensed spectrum. We have introduced the network architecture for LTE/LTE-A small cells to explore the unlicensed spectrum used by Wi-Fi systems. The ABS mechanism and an interference avoidance scheme have been proposed to mitigate the interference between Wi-Fi and LTE/LTE-A systems when both transmitting in the same unlicensed spectrum. Simulation results have shown that with the proper use of ABS mechanism and interference avoidance schemes, heterogeneous and small cell networks can improve their capacity by using the unlicensed spectrum used by Wi-Fi systems without affecting the performance of Wi-Fi.

References

- Almeida E, Cavalante A, Paiva R, Chaves F, Abinader F, Vieira R, Choudhury S, Tuomaala E, Doppler K (2013) Enabling LTE/Wi-Fi coexistence by LTE blank subframe allocation. In: IEEE international conference on communications (ICC), pp 5083–5088
- Bennis M, Simsek M, Czystlik A, Saad W, Valentin S, Debbah M (2013) When cellular meets WiFi in wireless small cell networks. *IEEE Commun Mag* 51(6):44–50
- Charitos M, Kalivas G (2013) Heterogeneous hybrid vehicular WiMAX-Wi-Fi network for intunnel surveillance implementations. In: IEEE international conference on communications (ICC), Budapest, pp 6386–6390
- Chu X, Lopez-Perez D, Yang Y, Gunnarsson F (eds) (2013) Heterogeneous cellular networks – theory, simulation and deployment. Cambridge University Press, London, pp 1–494. ISBN-13: 9781107023093
- Dimatteo S, Pan H, Han B, Li VOK (2011) Cellular traffic offloading through WiFi networks. In: IEEE 8th international conference on mobile Ad Hoc and sensor systems (MASS), Wuhan, pp 192–201
- Haenggi M (2005) On distances in uniformly random networks. *IEEE Trans Info Theory* 51(10):3584–3586
- Hajmohammad S, Elbiaze H (2013) Unlicensed spectrum splitting between Femtocell and Wi-Fi. In: IEEE international conference on communications (ICC), Budapest, pp 1883–1888
- Huawei, etc. (2013) Discussion paper on unlicensed spectrum integration to IMT systems, 3GPP RAN 62 RP-131723
- Hwang I, Song B, Soliman SS (2013) A holistic view on hyper-dense heterogeneous and small cell networks. *IEEE Commun Mag* 51(6):20–27
- Liang YS, Chung WH, Ni GK, Chen Y, Zhang H, Kuo SY (2012) Resource allocation with interference avoidance in OFDMA femtocell networks. *IEEE Trans Veh Technol* 61(5):2243–2255
- Nakamura T, Nagata S, Benjebbour A, Kishiyama Y, Tang H, Shen X, Yang N, Li N (2013) Trends in small cell enhancements in LTE advanced. *IEEE Commun Mag* 51(2):98–105
- Sooyoung H, Taejoon K, Love DJ, Krogmeier JV, Thomas TA, Ghosh A (2013) Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans Commun* 61(10):4391–4403
- Turyagyenda C, O'Farrell T, Guo W (2012) Energy efficient coordinated radio resource management: a two player sequential game modelling for the long-term evolution downlink. *IET Commun* 6(14):2239–2249
- Wang S, Guo W, McDonnell M (2014) Downlink interference estimation without feedback for heterogeneous network interference avoidance. In: IEEE international conference on telecommunications (ICT), Cairo, pp 1–6
- Yuan W, Wang X, Linnartz J, Niemegeers I (2010) Experimental validation of a coexistence model of IEEE 802.15.4 and IEEE 802.11 b/g networks. *Int J Distribut Sens Netw* 2010:1–6
- Zhang H, Jiang C, Beaulieu NC, Chu X, Wen X, Tao M (2014) Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services. *IEEE Trans Commun* 62(7):2366–2377
- Zhang H, Chu X, Guo W, Wang S (2015) Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum. *IEEE Commun Mag* 53(3):158–164
- Zhang H, Jiang C, Mao X, Chen H-H (2016) Interference-limit resource optimization in cognitive femtocells with fairness and imperfect spectrum sensing. *IEEE Trans Veh Technol* 65(3):1761–1771
- Zhang H, Huang S, Jiang C, Long K, Leung VCM, Poor HV (2017a) Energy efficient user association and power allocation in millimeter wave based ultra dense networks with energy harvesting base stations. *IEEE J Sel Areas Commun* 35(9):1936–1947

- Zhang H, Qiu Y, Chu X, Long K, Leung VCM (2017b) Fog radio access networks: mobility management, interference mitigation and resource optimization. *IEEE Wirel Commun* 24(6):120–127
- Zhang H, Qiu Y, Long K, Karagiannidis GK, Wang X, Nallanathan A (2018) Resource allocation in NOMA based fog radio access networks. *IEEE Wirel Commun* 25(3):110–115

Heterogeneous Wireless Access Medium with Cognitive Radio

- [Cognitive Heterogeneous Networks](#)

Heterogeneous Wireless Networks Based on Cognitive Radio

- [Cognitive Heterogeneous Networks](#)

HetNet

- [Mobile Edge Caching in HetNets](#)

Hidden and Exposed Terminal Problems

Zijun Gong and Fan Jiang
Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

Synonyms

[Exposed node problem](#); [Hidden node problem](#)

Definition

The hidden node problem refers to the situation where two nodes try to send packets to a third node simultaneously because they don't know the existence of each other. The exposed node problem refers to the situation where a node refrains from transmitting because a neighbor node is sending packets.

Historical Background

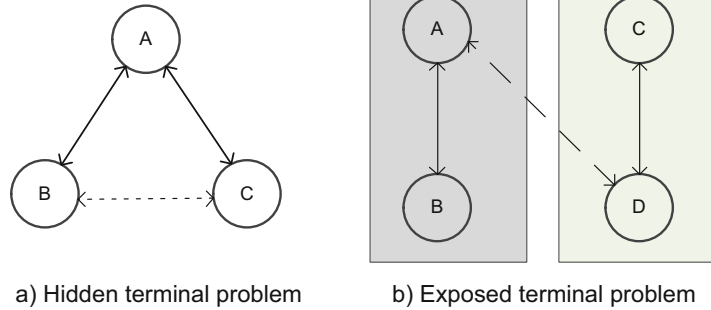
The well-known hidden terminal and exposed terminal problems severely reduce network throughput in wireless ad hoc networks, as shown in Fig. 1.

The hidden terminal problem is illustrated in subfigure (a). As we can see, node A can hear node B and node C, while node B and node C cannot hear each other. Assume node B is sending data to node A and node C also has data to send. In this case, node C doesn't know that node B is transmitting and will start data transmission to node A. As a result, packets from node B and C will collide at node A. The exposed terminal problem is depicted in subfigure (b). Node A and B are deployed in Area I, while node C and D are deployed in Area II. Assume A can hear B and C can hear D. Signal from A cannot be heard by C, but can be detected by D. In this case, when A is sending packets to B, D can safely send packets to C at the same time without collision. However, D will refrain from doing this, because D mistakenly thinks that the channel is occupied.

As we can see, both the hidden terminal problem and the exposed node problem will reduce network throughput for different reasons. For the hidden terminal problem, performance loss is caused by the retransmissions after collisions. On the other hand, the exposed terminal problem is not efficient because some communications that can be conducted simultaneously are blocked. Many different protocols have been proposed to overcome these two issues. The basic idea is to add a control channel, on which communi-

Hidden and Exposed Terminal Problems,

Fig. 1 The hidden terminal and exposed terminal problems



cations among different nodes can be coordinated. Here, we will briefly review some of these protocols.

The busy tone multiple access (BTMA) protocol was proposed in 1975 to tackle these issues (Tobagi and Kleinrock 1975). In BTMA, the channel between transmitter and receiver is divided into a data channel and a control channel. These two channels lie on different frequencies. When a transmitter wants to send packets to a receiver, it will first sense the control channel. If a busy tone is sensed, the transmitter will back off randomly and transmit the packets later. On the other, it will start transmission and put a busy tone on the control channel simultaneously. When the receiver senses the data on its data channel, it will also put a busy tone on its control channel. This protocol successfully reduced collision probability. However, the bandwidth utilization is very low.

The MACA (Multiple Access with Collision Avoidance) scheme is another scheme (Karn 1990). Before sending packets, the sender will send a *request to send* (RTS) message to the corresponding receiver. If the receiver is currently ready to receive data, it will reply a *clear to send* (CTS) message. Then, the sender can safely send data to its intended receiver. Reference (Bharghavan et al. 1994) presented a variant of MACA.

To improve bandwidth utilization, the dual busy tone multiple access (DBTMA) protocol was proposed in Deng and Haas (1998). Similar to BTMA, it has both data and control channels. However, there are four types of signals on the control channel: the request to send packet, the

clear to send packet, the busy tone for transmitting (BTT), and the busy tone for receiving (BTR). Whenever a node wants to transmit packets, it will first sense the control channel. If there is a BTR, it means that a neighbor node is receiving packets, and the transmitter needs to wait. Otherwise, the transmitter will send a RTS packet. Upon receiving the RTS packet, the receiver will first sense the control channel. If BTT is detected, it means that a neighbor node is transmitting, and the receiver cannot receive packets at this moment. On the other hand, the receiver will send the transmitter a CTS packet and put a BTR on control channel. Upon receiving the CTS packet, the transmitter will put a BTT on control channel and start transmission. After all the packets are transmitted, both sides will turn off the busy tones. Compared with the other RTS/CTS-based protocols, the DBTMA shows better network utilization.

Another variant of BTMA is the receiver-initiated tone multiple access protocol (RI-BTMA) (Wu and Li 1987). In this protocol, both control channel and data channel are divided into slots of the same length. When a node wants to transmit something, it first senses the control channel. If busy tone is detected, it will wait for the next free slot. Then, a preamble is transmitted on data channel, which contains the identification of the intended destination node. Upon receiving the preamble, the destination node will put a busy tone on the control channel, until all packets have been successfully received. After transmission, both receiver and transmitter will cancel the busy tone. On the other hand, if the preamble is not successfully received by the intended receiver,

the sender will not receive acknowledgment from the receiver. In this case, the sender will wait for the next free slot until communication is built up. The RI-BTMA protocols can be divided into two types: the basic one and the controlled one. For the basic type, nodes cannot buffer data on collisions. As a result, collided packets cannot be retransmitted. Besides, when network load is intense, this protocol is not stable, because nodes do not have queuing system for incoming packets. To solve these problems, the controlled protocol is proposed. Nodes can buffer the failed packets and retransmit them. Once the buffers are not empty, the node will work in backlogged mode. In this case, the node will first retransmit the failed packets and buffer the incoming packets.

Foundations

This part focus on analyzing how the reviewed protocols overcome the hidden and exposed terminal issues. Besides their advantages and limitations will also be discussed. Moreover, we will briefly introduce the IEEE 802.11 RTS/CTS mechanism.

Adopting the example in Fig. 1, for BTMA, when node B is sending packets to node A, node A will put a busy tone on the control channel. When node C also wants to send packets to node A, it will first sense the control channel. Due to the busy tone, node C immediately understands that node A is receiving packets from some other node. Therefore, it will wait until the busy tone is cancelled. As a result, the hidden node issue is resolved. For the exposed node issue, when node A is transmitting packets to node B, node B will put a busy tone on control channel. When node D wants to send packet node C, it will first sense the control channel. Because node B is out of the communication range of node D, the busy tone cannot be heard by node D. Therefore, although node D hears node A is transmitting, it knows that the receiver is not in its range. Node D can safely and confidently start its transmission to node C. This protocol has very low complexity and can be easily implemented. However, it cannot be used

in highly populated networks because of its low efficiency.

The MACA protocol works in a different way. Still adopting the example in Fig. 1, when node B is transmitting to node A, node A puts a busy tone on control channel. Therefore, node C knows that A is receiving packets from other nodes, which means the hidden terminal problem is solved. On the other hand, when node A is transmitting to node B, B will put a busy tone on control channel. Node D overhears node A, but it cannot hear the busy tone from node B. As a result, node D realizes that the receiver (node B) is not in its communication range, and it is safe to start the transmission to node C.

The IEEE 802.11 RTS/CTS mechanism is a variant of the MACA protocol. In this mechanism, when a node wants to transmit packets to a destination node, it will first send RTS. When the destination node replies with CTS, the sender will transmit a short Data-Sending packet, which contains the information about the length of the data packet to be transmitted. After this, the data packet can be transmitted, and the receiver will reply with ACK. It should be noted that the sender can only transmit one packet in every round for fairness. After every transmission, the sender waits for a random time and competes for medium access with other nodes. If a node overhears RTS, it means some neighbor node is waiting for the CTS message. Therefore, this node needs to wait until it overhears CTS or for a certain time so as to guarantee the un-interfered reception of the CTS of the neighbor node. On the other hand, if a node overhears CTS, it means that some neighbor node is trying to receive data. Therefore, this node needs to wait until the transmission is over, as indicated by ACK. If a node is in the vicinity of the sender but not the receiver, it can hear the RTS and DS, which indicate the successful exchange of RTS/CTS. Therefore, it needs to wait for the estimated transmission time plus a random back off.

This protocol is not fair in some scenarios. For example, when the receiver is in the communication range of an active node receiving packets from another node, the incoming RTS will be totally blocked. To be specific, the sender

sends RTS to the receiver, but the receiver knows that someone in its neighbor is receiving packets, and it cannot reply. The sender assumes the RTS failed and backs off randomly. If the active neighbor of the receiver keeps receiving packets, the communication between the sender and receiver can be blocked for a long time, which is unfair. To overcome this problem, the request for request to send (RRTS) is introduced into the protocol. When the receiver receives RTS and cannot reply immediately, it will wait. When it overhears ACK, it will immediately send a RRTS to the sender, and the sender will reply with RTS. Then, the receiver sends CTS, and the sender can safely commence data transmission. In this case, the previously active node overhears the RRTS, and it will wait for two time slots. After this, if it hears CTS from the target receiver, that means the communication is successfully built up, and it has to wait until it is over.

It should be noted that the exposed terminal problem is not completely solved in IEEE 802.11 RTS/CTS. For example, when the sender transmits a RTS to the receiver, if the receiver is in the transmission range of another active node sending data, the RTS will collide with the data packets. As a result, the RTS is not successfully received by the receiver due to co-channel interference, and the sender will wait for the CTS until time out. Then, the sender assumes that the failure of RTS was caused by collision, and it will back off for a random time and try again. Generally, data packets are much longer than control packets, and the sender will keep trying until the RTS is successfully transmitted between transmissions of the active node. This leads to low efficiency and unfairness and is one of the flaws of this protocol.

Key Applications

Multiple access collision avoidance is key technology for any communication protocols in ad hoc networks.

Cross-References

- [Ad Hoc Networks](#)
- [Multiple Access Techniques](#)

References

- Bharghavan V et al (1994) MACAW: a media access protocol for wireless LANs. In: Proceedings of ACM SIGCOMM 1995, London
- Deng J, Haas ZJ (1998) Dual busy tone multiple access (DBTMA): a new medium access control for packet radio networks. Universal Personal Communications, 1998. In: IEEE 1998 international conference on ICUPC '98, Florence, vol 2
- Karn P (1990) MACA – a new channel access method for packet radio. In: Proceedings of ARRL/CRRL amateur radio computer networking conference 1990, Ontario
- Tobagi F, Kleinrock L (1975) Packet switching in radio channels: part II – the hidden terminal problem in carrier sense multiple-access and the busy-tone solution. IEEE Trans Commun 23(12):1417–1433
- Wu C, Li V (1987) Receiver-initiated busy-tone multiple access in packet radio networks. In: Proceedings of the ACM workshop on frontiers in computer communications technology (SIGCOMM '87), New York

Hidden Node Problem

- [Hidden and Exposed Terminal Problems](#)

High-Frequency MIMO

- [Millimeter-Wave \(mmWave\) Multiple-Input Multiple-Output \(MIMO\) Technique](#)

HTTP Adaptive Streaming

- [QoE Measurement and Assessment of Video Streaming](#)

Hybrid Beamforming

► Hybrid Precoding

Hybrid Precoding

Yue Xiao, You Li, and Yanping Xiao
National Key Laboratory of Science and
Technology on Communications, University of
Electronic Science and Technology of China,
Chengdu, China

Synonyms

Hybrid beamforming

Definition

Hybrid precoding is one of the precoding techniques to support noninterference multi-stream transmission in massive multiple-input multiple-output (MIMO) wireless communications. More explicitly, hybrid precoding is a combination of the analog precoder in the radio-frequency (RF) front end, together with the digital precoder in the baseband. A major advantage of hybrid precoding is that it consumes smaller number of up/down conversion chains than the conventional full digital precoder for MIMO systems. In hybrid precoding, the introduced analog precoder is a network of analog phase shifters, which control the phase of the signal on each transmit antenna. Therefore, hybrid precoding is capable of providing a significant precoding gain but with low hardware cost and low power consumption.

Hybrid precoding has two main structures as shown in Fig. 1 (Han et al. 2015; Molisch et al. 2017; Huang et al. 2018). In the full-complexity structure as shown in Fig. 1a, the digital signal from each transceiver can be delivered to any antenna after analog weighting. Thus, each analog precoder output can be a linear combination

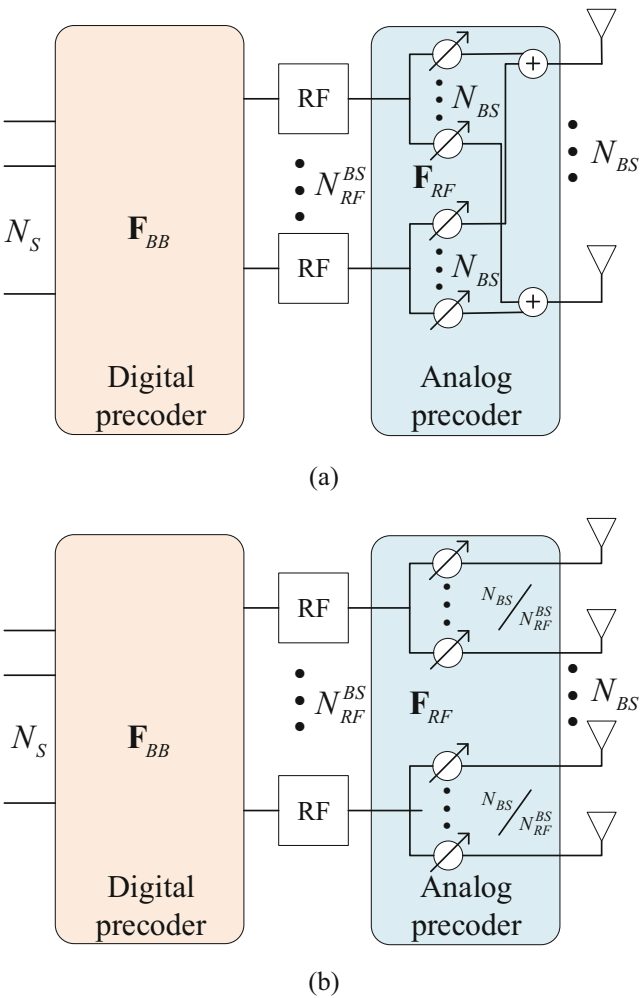
of all the RF signals with a relatively high complexity. By appropriately designing the analog and digital precoders, the full-complexity structure is capable of reaching the capacity upper bound, i.e., the capacity of full digital precoder. By contrast, in the reduced-complexity structure as shown in Fig. 1b, the digital signal from each transceiver can only be delivered to a fixed number of antennas, which causes performance loss compared with the full-complexity structure. Thus, each analog precoder output can be a phase shift version of its corresponding RF signal, and the design for the analog precoder is much easier than that of its full-complexity counterpart.

Hybrid precoding scheme can also be divided into two types as full feedback and limited feedback. In the full feedback hybrid precoding scheme, the transmitter can acquire the full channel state information (CSI). Taking advantage of the full CSI, the transmitter can appropriately arrange the analog and digital precoders and provide the same performance as the full digital precoder (Bogale et al. 2016). However, since it is difficult to feedback all the channel information in massive MIMO, full CSI is hard to be directly obtained at the transmit side. Thus, the limited feedback hybrid precoding scheme was proposed (Alkhateeb et al. 2015; Liu et al. 2018). The limited feedback hybrid precoding is generally divided into three stages. At the first stage, the transmitter broadcasts all the codebook vectors in turn. At the second stage, the receiver feedbacks the indices of the selected codebook vectors based on designed criterion. At the third stage, the transmitter uses the selected codebook vectors to generate the analog and digital precoding books.

Historical Background

Massive MIMO has attracted lots of research interests due to its outstanding performance toward the downlink transmission of future wireless communications (Xiao et al. 2017; Xue et al. 2016; Cui et al. 2018). However, the large number of antenna elements in massive MIMO also poses major challenges. Most

Hybrid Precoding, Fig. 1
Two structures of hybrid precoding



important of all, a large number of RF chains are introduced, which increases the cost and energy consumption.

A promising solution to the above problem lies in the concept of hybrid transceivers, which use a combination of the analog precoder in the RF front end, together with the digital precoder in the baseband, connected to the RF with a smaller number of up/down conversion chains. Hybrid precoding was first introduced and analyzed in the mid-2000s in Zhang et al. (2005) and Sudarshan (2006). It is motivated by the fact that the number of up/down conversion chains is only lower-limited by the number of data streams, while the precoding gain and diversity order is

given by the number of antenna elements if suitable RF precoding is done.

Key Applications

The key applications of hybrid precoding include massive MIMO and cellular and millimeter wave communications.

Cross-References

- [Millimeter Wave Beam Training and Tracking](#)
- [Millimeter Wave Channel Measure](#)
- [Millimeter-Wave Massive MIMO](#)

References

- Alkhateeb A, Leus G, Heath RW (2015) Limited feedback hybrid precoding for multi-user millimeter wave systems. *IEEE Trans Wirel Commun* 14(11):1536–1276
- Bogale TE, Le LB, Haghighat A, Vandendorpe L (2016) On the number of RF chains and phase shifters, and scheduling design with hybrid Analog-digital beamforming. *IEEE Trans Wirel Commun* 15(5):3311–3326
- Cui Y, Fang X, Fang Y, Xiao M (To Appear, 2018) Optimal non-uniform steady mmwave beamforming for high speed railway. *IEEE Trans Veh Technol*. <https://doi.org/10.1109/TVT.2018.2796621>
- Han S, Chih-Lin I, Xu Z, Rowell C (2015) Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G. *IEEE Commun Mag* 53(1):186–194
- Huang Y, Zhang J, Xiao M (To Appear, 2018) Constant envelope hybrid precoding for directional millimeter-wave communications. *IEEE J Sel Areas Commun JSAC Series Phys Layer Security 5G Wirel Netw* <https://doi.org/10.1109/JSAC.2018.2825820>
- Liu Y, Fang X, Xiao M, Mumtaz S (To Appear, 2018) Decentralized beam pair selection in multi-beam millimeter-wave networks. *IEEE Trans Commun*. <https://doi.org/10.1109/TCOMM.2018.2800756>
- Molisch AF, Ratnam VV, Han S, Li Z, Nguyen SLH, Li L, Haneda K (2017) Hybrid beamforming for massive MIMO: a survey. *IEEE Commun Mag* 55(9):134–141
- Sudarshan P (2006) Channel statistics-based RF pre-processing with antenna selection. *IEEE Trans Wirel Commun* 5(12):3501–3511
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis G, Bjornson E, Yang K, Chih-lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun* 35(9):1909–1935
- Xue Q, Fang X, Xiao M, Li Y (2016) Multi-user millimeter wave communications with nonorthogonal beams. *IEEE Trans Veh Technol* 66(7):5675–5688
- Zhang X, Molisch A, Kung S-Y (2005) Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection. *IEEE Trans Sign Proces* 53(11):4091–4103

Hybrid System

- [Smart Device-to-Device Communication: Communication Establishment and Maintenance](#)

Hyper Cellular Network: Control-Traffic Decoupled Radio Access Network Architecture

Sheng Zhou and Zhisheng Niu
Tsinghua University, Beijing, China

Synonyms

[Control-data separation](#); [Dual connectivity](#)

Definitions

Hyper cellular network is a radio access network architecture based on the separation of control coverage and traffic coverage, enabling agile coverage, elastic access, and adaptive service.

Background

A base station (BS) in traditional radio access networks (RANs) handles both the signaling control and data transmission for users in its coverage. The tight coupling of control and data transmission on a single entity makes it hard to provide flexible and energy-efficient services for RANs. To meet the increasing needs for green and elastic wireless access, new air interface architectures of cellular networks that feature signaling and data separation have been developed. Hyper cellular network architecture is one of them that aims at flexible and efficient control of small cells for throughput boosting and BS sleeping-based energy saving (NIU et al. 2012).

To realize the decoupled control-traffic coverage, separation schemes for the air interface have been recently developed in Xu et al. (2013), Zhao et al. (2015) and Jha et al. (2014). To make cell coverage more adaptive to traffic dynamics, some control signaling functions should be decoupled from the data functions so that the data traffic is served on demand, while the control plane is

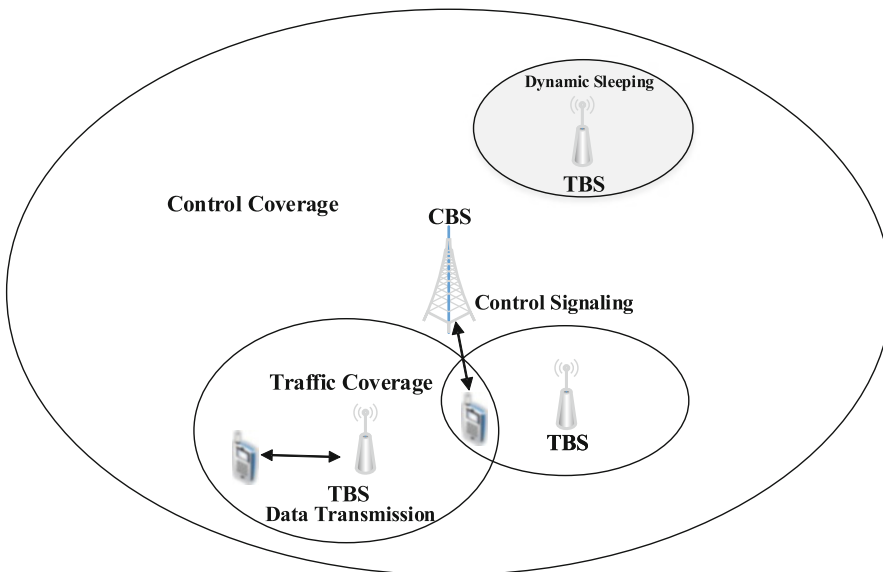
always on to guarantee basic coverage. Based on functionality separation (Xu et al. 2013), a separation scheme is proposed in Zhao et al. (2015), which designed the separation of logical channels and signaling protocols in the hyper cellular network. Besides, the idea of decoupled air interface in the cellular standard is known as dual connectivity in Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) Release 12 (3GPP 2013). In dual connectivity, the separation is specified in the aspects of the control plane and user plane (Jha et al. 2014).

Foundations

As illustrated in Fig. 1, the hyper cellular network decouples the air interface of RAN by separating the control coverage from the traffic coverage with two types of base stations (BSs). Control BSs (CBSs) provide the control coverage, while traffic BSs (TBSs) provide the traffic coverage. Typically CBSs have large coverage areas, within which TBSs are deployed. CBSs provide the network control for underlying TBSs and mobile users. In particular, they grant the network

access to mobile users. To guarantee a basic level of data service and to cope with high mobility users, CBSs can also be used to provide low-rate data services such as voice calls to the user equipment (UE). Unlike CBSs, TBSs are only responsible for high-rate data services. Specifically, there can be different types of TBSs to support various classes of high-rate and low-mobility data services. Through the separation, TBSs can be much more simplified than conventional BSs and can simply go to sleep whenever there is no or very few traffic within their coverages. Through this kind of sleeping control of the traffic coverage, a large amount of energy consumption can be saved because the major part of energy consumption of BSs is the basic power (baseband processing unit, amplifier, fans, etc.), which can only be saved by turning the whole BS off.

Under the separation architecture, CBSs provide centralized controls and enable dynamic operations of TBSs. CBSs and TBSs serve mobile users collaboratively. When a user equipment (UE) is powered on, it searches for nearby CBSs and registers to the network through a CBS. CBSs gather network state information such as UE locations and traffic load from UEs and TBSs



Hyper Cellular Network: Control-Traffic Decoupled Radio Access Network Architecture, Fig. 1 A sketch for the separation of control coverage and traffic coverage in hyper cellular network

underlying its coverage and thus hold a global view of the network. When high-rate data services are required, the UE sends requests to the CBS, and the CBS dispatches one or more TBSs for the requested service afterward. When the network traffic load is light, TBSs can be turned off under the command of the CBS to reduce the energy consumption of the network; in the meantime the CBS still preserves the network coverage.

To realize the decoupled architecture, network functionalities and logical channels need to be separated for different types of BSs. In hyper cellular network, CBSs are in charge of synchronization, broadcast of system information, paging, and low-rate data transmission, while TBSs are only responsible for high-rate data transmissions. The separation scheme for the logical channels is shown in Table 1, considering the corresponding network functionalities. BCCH (broadcast control channel, BCH in GSM/GPRS) and CCCH (common control channel, including PCCH for paging in UMTS and LTE) are reserved by CBSs since they are in charge of network functionalities including synchronization, paging, and broadcast of system information. CTCH (common traffic channel, TCH in GSM/GPRS, MTCH and MCCH for multicast in LTE) is also left at the CBS side to take care of low-rate data transmissions. At the TBS side, DTCH (dedicated traffic channel) and DCCH (dedicated control channel) are supported, and they are responsible for high-rate data transmissions. In the case of GSM/GPRS, PDTCH (packet data traffic channel) and PACCH (packet associated control

channel) are exploited at TBSs for packet data transmissions and associated controls.

In RANs, radio resource control (RRC) connection setup process is required between the random access and the setup of data services. Specifically in hyper cellular network, for high-rate data services, UE has access to the CBS first and then builds connection to the TBS. To this end, the RRC connection setup process needs to be modified, plus the signaling interaction between different BSs. As shown in Fig. 2, UE first sends RRC connection request message to the CBS. The CBS then forwards the message to the TBS according to a given BS dispatching algorithm. After that, the TBS sends RRC connection setup messages with channel information to the CBS, forwarding to the UE. Finally, the UE sends RRC connection setup complete message and builds the initial connection to the TBS. After the RRC connection is set up, the following procedures are similar to those in current RANs. A high-level flowchart for data services in hyper cellular network is shown in Fig. 3.

Key Applications

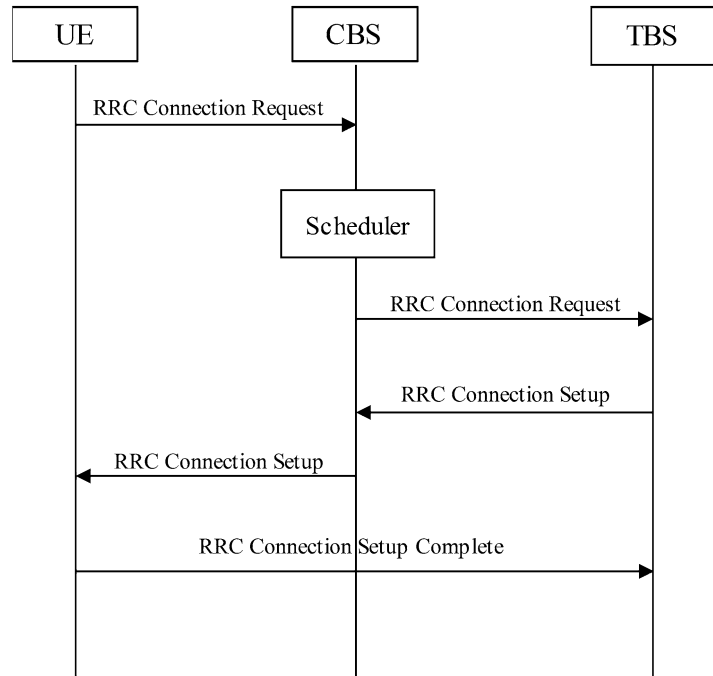
The decoupled structure of RAN enables many new applications and technologies for green and elastic access in wireless communications. The following are two typical applications of hyper cellular network:

- Load balancing. In hyper cellular network, the CBS can gather the information of all UEs and TBSs within its coverage and thus have a global view of the network. Based on that, the CBS can dispatch multiple TBS for load balancing, which can potentially deliver better services in addition to improving radio resource utilization and reducing the risk of system failure. To implement load balancing, the active TBSs periodically send the load information to the CBS through the backhaul link. When a new channel request arrives, the CBS decides which TBS is going to provide the service based on the load information.

Hyper Cellular Network: Control-Traffic Decoupled Radio Access Network Architecture, Table 1 Logical channel separation in hyper cellular network

Logical channel			Separation	
GSM/GPRS	UMTS	LTE	CBS	TBS
BCH	BCCH	BCCH	✓	
CCCH	CCCH, PCCH	CCCH, PCCH	✓	
PACCH	DCCH	DCCH		✓
TCH	CTCH	MTCH, MCCH	✓	
PDTCH	DTCH	DTCH		✓

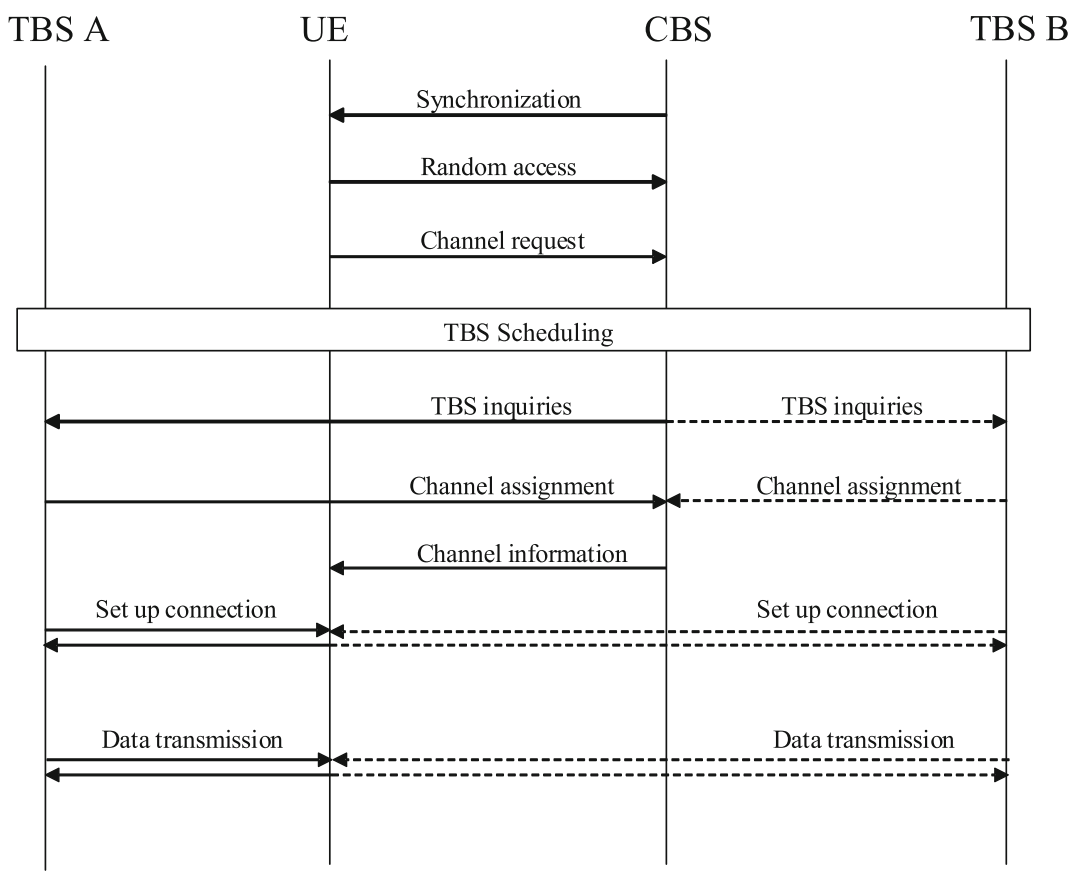
**Hyper Cellular Network:
Control-Traffic
Decoupled Radio Access
Network Architecture,
Fig. 2** RRC connection
setup process in hyper
cellular network



- BS sleeping-based energy saving. With the decoupled air interface, the CBS can provide an “always-on” control coverage, which enables the flexible and dynamic sleeping of TBSs without generating unwanted coverage holes. Because the CBS holds the global network information, it makes the decisions which TBSs are going to sleep based on a given BS sleeping algorithm. Researchers in Zhang et al. (2014) and Guo et al. (2016) have developed several BS sleeping algorithms and shown the benefits of BS sleeping in energy saving. To control the dynamic sleeping of TBSs, the CBS can periodically send indication messages to each TBS, based on which the TBS can decide its work state, i.e., active or sleep. The messages are transmitted through the backhaul links between the CBS and the TBSs.
- Integrating the hyper cellular concept with cloud-based RAN architecture and software-defined network technology. Cloud RAN (C-RAN) (Chih-Lin et al. 2014) consolidates baseband units (BBUs) of BSs to a centralized computing cloud while only leaving remote radio units at the network edge. With virtualization, virtual base stations (VBSs) can be realized therein for flexible RAN configurations and operations. SoftRAN (Gudipati et al. 2013) aims to enable software-defined RANs (SDRANs) by introducing a logically centralized controller and distributed radio elements. With the centralized controller, SDRAN simplifies network management by software programming. CONCERT (Liu et al. 2014) builds a converged cloud and cellular system based on control-data decoupling. Some initial efforts on realizing hyper cellular network via a software-defined approach in a cloud-based infrastructure can be found in Zhou et al. (2016), where a software-defined hyper cellular architecture (SDHCA) is proposed.
- Learning-based state control for TBSs in hyper cellular network. The optimal working

Future Directions

There are some future directions beyond the separation of the air interface in hyper cellular networks, two of which are listed as follows:



Hyper Cellular Network: Control-Traffic Decoupled Radio Access Network Architecture, Fig. 3 Flow chart of high-rate data services in hyper cellular network

states of TBSs depend on many factors, among which the channel state information (CSI) and traffic load variations are the keys. There is a fundamental difficulty to get the CSI of sleeping TBSs as no pilot is sent from them. To estimate the CSI of sleeping TBSs in order to assist the decision of potentially turning on sleeping TBSs, users can exploit the correlation among the CSI of sleeping TBSs and active TBSs or the “always-on” CBS, via learning mechanisms (Liu et al. 2015). Initial efforts of learning the CSI of sleeping TBSs from CBS equipped with multiple antennas can be found in Chen et al. (2017). In terms of traffic variations, conventional mechanism on TBS on/off

switching mostly rely on simple statistical models, e.g., Poisson models. However, real traffic largely deviates from these ideal models. To this end, model-free TBS control mechanisms based on reinforcement learning are yet to be explored.

Cross-References

- [Energy Efficiency of Cellular Networks](#)
- [Future Wireless Network Architecture and Network Slicing](#)
- [Heterogeneous Small Cell Networks](#)
- [5G Wireless](#)

References

- 3GPP (2013) 3GPP TR 36.842 v12.0.0: study on small cell enhancements for e-utra and e-utran-higher layer aspects. Technical report, 3rd generation partnership project. http://www.3gpp.org/ftp/Specs/archive/36_series/36.842/36842-c00.zip
- Chen S, Jiang Z, Liu J, Vannithamby R, Zhou S, Niu Z, Wu Y (2017) Remote channel inference for beamforming in ultra-dense hyper-cellular network. In: 2017 IEEE global communications conference (GLOBECOM)
- Chih-Lin I, Huang J, Duan R, Cui C, Jiang JX, Li L (2014) Recent progress on c-ran centralization and cloudification. *IEEE Access* 2:1030–1039
- Gudipati A, Perry D, Li LE, Katti S (2013) Softran: software defined radio access network. In: Proceedings of the second ACM SIGCOMM workshop on hot topics in software defined networking. ACM, pp 25–30
- Guo X, Niu Z, Zhou S, Kumar P (2016) Delay-constrained energy-optimal base station sleeping control. *IEEE J Sel Areas Commun* 34(5):1073–1085
- Jha SC, Sivanesan K, Vannithamby R, Koc AT (2014) Dual connectivity in LTE small cell networks. In: 2014 Globecom workshops (GC Wkshps). IEEE, pp 1205–1210
- Liu J, Zhao T, Zhou S, Cheng Y, Niu Z (2014) Concert: a cloud-based architecture for next-generation cellular systems. *IEEE Wirel Commun* 21(6):14–22
- Liu J, Deng R, Zhou S, Niu Z (2015) Seeing the unobservable: channel learning for wireless communication networks. In: Proceedings of IEEE global communication conference
- Niu Z, Zhou S, Zhou S, Zhong X, Wang J (2012) Energy efficiency and resource optimized hyper-cellular mobile communication system architecture and its technical challenges. *SCIENTIA SINICA Informationis* 42(10):1191–1203
- Xu X, He G, Zhang S, Chen Y, Xu S (2013) On functionality separation for green mobile networks: concept study over LTE. *IEEE Commun Mag* 51(5):82–90
- Zhang S, Wu J, Gong J, Zhou S, Niu Z (2014) Energy-optimal probabilistic base station sleeping under a separation network architecture. In: 2014 IEEE global communications conference (GLOBECOM). IEEE, pp 4239–4244
- Zhao T, Wang L, Zheng X, Zhou S, Niu Z (2015) Hycell: enabling green base station operations in software-defined radio access networks. In: 2015 IEEE international conference on communication workshop (ICCW). IEEE, pp 2868–2873
- Zhou S, Zhao T, Niu Z, Zhou S (2016) Software-defined hyper-cellular architecture for green and elastic wireless access. *IEEE Commun Mag* 54(1):12–19

Hyper MIMO Channel Estimation

► Massive MIMO Channel Estimation



ID/Locator Separation-Based Mobility Management

- [Mobility Management with ID/Locator Separation](#)

ID/Locator Split-Based Mobility Management

- [Mobility Management with ID/Locator Separation](#)

Identifier Network

- [Smart Identifier Network](#)

IEEE 1609 Standards Set

- [Wireless Access in Vehicular Environment](#)

IEEE 802.11

- [Authentication](#)

IEEE 802.1x

- [Authentication](#)

IEEE Std 802.11p

- [Wireless Access in Vehicular Environment](#)

Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks

Lei Mo, Angeliki Kritikakou, and Olivier Sentieys
Univ Rennes, INRIA, CNRS, IRISA, Rennes, Cedex, France

Synonyms

[Imprecise computations](#); [Multicore architectures](#); [Task mapping](#)

Definitions

Imprecise Computation (IC) model (Liu et al. 1994) considers that a task can be logically

decomposed into a *mandatory* subtask and an *optional* subtask. The mandatory subtask should be completed before the task's deadline in order to generate the minimum acceptable Quality of Service (QoS). The optional subtask is to be executed after the mandatory subtask, and still before the deadline, if there are resources in the system that are free to execute the optional subtask. The longer the optional subtask is executed, the better is the QoS of the obtained result.

Historical Background

Wireless Sensor Networks (WSNs) consist of a large number of wireless sensor nodes that have the capability to take various measurements of their environment. These measurements could be temperature, sound, vibration, pressure, motion, etc. *Energy efficiency* and *real-time execution* are critical and challenging issues for the design of WSNs systems (Mo et al. 2018a). This is because:

1. Most of the sensor nodes have energy constraints, as they are powered by a battery that has a limited energy budget.
2. Real-time responsiveness is required by many WSN applications, e.g., mobile target tracking. Missing of the task deadline can cause serious results.

In the past, many research works have focused on optimizing the energy efficiency and real-time execution of applications on WSNs, e.g., data aggregation (Dhasian and Balasubramanian 2013), topology control (Esch 2013), sink mobility (Silva et al. 2014), and delay-aware routing (Lai et al. 2018). These methods improved significantly the energy efficiency and real-time execution of the WSNs. However, further improvements exist:

1. Most of the past works focused on the software optimization methods without including the hardware optimization strategies.

2. Most of the past works focused on reducing the energy consumption of the wireless transmission but ignored the energy consumption of the processor modules.

However, processing and transmitting large amount of sensed data in real-time applications exceeds the capabilities of traditional WSNs. Since *single-core* embedded sensor nodes will soon be unable to meet the increasing requirements of data intensive applications, e.g., Wireless Multimedia Sensor Networks (WMSNs), the next generation sensor nodes must possess enhanced computation and communication capabilities, i.e., *multi-core* embedded sensor nodes (Munir et al. 2014). For instance, the fire detection WMSN node MiLive-v2 in (Liu et al. 2018) is equipped with three types of processors:

1. A 8-bit low-power AVR processor (ATmega128rfa1), which can be used to run simple tasks.
2. A 32-bit powerful ARM processor (ARM1176JZF), which can be used to run more complicated tasks.
3. A Digital Signal Processor (DSP) unit, which can be specially used to perform image processing.

The requirements of WSNs are usually high-performance computation, low-energy consumption, and low-task execution delay. The efficient mapping of the tasks on multi-core sensor nodes (Liu et al. 2018) is required to balance these *contradictory* requirements. It can simultaneously optimize *task allocation* (on which processor each task is executed) and *task scheduling* (when each task starts and ends its execution). On the one hand, in order to map a set of applications on multi-core system, the applications need to be partitioned (parallelized), whenever possible, into multiple tasks that can be executed concurrently on different processors. Therefore, task execution delay can be further reduced. On the other hand, multi-core systems have better energy savings than traditional single-core systems in two ways:

1. Multi-core embedded sensor nodes can extract the desired information from the sensed data and process this information. It leads to reduction in the data transmission volume sent to the sink node (base station). By replacing a large percentage of communication with in-network computation, the multi-core system realizes large energy savings that increase the sensor network's overall lifetime.
2. A multi-core embedded sensor node allows the computations to be split across multiple processors, while running each processor at a lower voltage and frequency, as compared to a single-core system, which results in energy savings. Utilizing a single-core embedded sensor node for information processing in data intensive applications demands the sensor node run at a high voltage and frequency in order to meet the application's real-time requirements. It will increase the power dissipation of the processor.

The tradeoff between energy consumption and performance has been extensively studied using the following methods:

1. *Dynamic Voltage and Frequency Scaling* (DVFS): By lowering the supply voltage, quadratic savings in dynamic energy consumption can be achieved, while the performance is approximately degraded in linear manner.
2. *Dynamic Power Management* (DPM): By exploring idle intervals of the processors and switching off a processor, when the idle interval is longer than a certain threshold, which is called break-even time.

Existing works that focus on energy-aware task mapping problem aim at minimizing the energy consumption under resource and application constraints.

1. When the voltage level is *discrete*, the task mapping problem is usually formulated as Integer Programming (IP), e.g., (Mahmood et al. 2017; Xu et al. 2012; Li and Wu 2015). To efficiently solve the IP problem, a hybrid

Genetic Algorithm (GA) is proposed in (Mahmood et al. 2017), a polynomial-time two-step heuristic is designed in (Xu et al. 2012), and the IP problem is relaxed to a Linear Programming (LP) in (Li and Wu 2015). Combining DVFS and DPM, a Mixed-Integer Linear Programming (MILP)-based task mapping problem is considered in (Chen et al. 2014) and the problem is solved by CPLEX solver.

2. When the voltage level is *continuous*, a convex task mapping problem is formulated in (Kong et al. 2011) and the problem can be solved by using polynomial-time methods. In (Leung et al. 2003), Mixed-Integer Non-Linear Programming (MINLP) is used to formulate the task mapping problem. The problem is relaxed to an MILP by linear approximation and solved by Branch and Bound (B&B) method.

As long as the resource and application constraints (e.g., task deadline) allow, the methods with DVFS or DPM achieve significant energy reductions compared to the basic methods, where no DVFS or DPM is applied.

Key Applications

In several application domains, an approximate – but still in time – result is preferable than an exact result obtained too late. For example, in audio and video streaming, frames with a lower quality are better than missing frames. In radar tracking, an estimation of target's location in time is better than an accurate location arriving too late. In control loops, an approximate result produced by a control scheme is more preferable as long as the controlled system, e.g., indoor temperature control system, remains stable. This type of applications can be modeled as *Imprecise Computation* (IC) tasks. The more cycles of the optional subtasks are executed, the higher is the generated QoS. An interesting question is how should the function of QoS be defined in order to represent realistic application areas:

1. A *linear* function (Yu et al. 2013a, b; Zhou et al. 2017; Wei et al. 2017; Mo et al. 2017, 2018b) models the case where the QoS of the overall system increases uniformly during the optional subtask execution.
2. A *concave* function (Aydin et al. 2001; Yu et al. 2008; Cortes et al. 2006; Tchamgoue et al. 2013) addresses the case where the increase of QoS exhibits a continuously nondecreasing rate as the optional subtask execution goes on.

Linear and general concave functions are considered as the most realistic and typical QoS representation in the literature, as they adequately capture the behavior of many application areas, such as image and speech processing, control engineering, and automatic target recognition (Aydin et al. 2000, 2001; Liu et al. 1994).

Foundations

The three constraints *energy*, *deadline*, and *QoS* play important roles in the QoS-aware task mapping. This is different from the energy-aware task mapping, where no exploration of the QoS improvement through the optional subtasks adjustment take place. The QoS-aware task mapping problem uses the IC task model and has a goal to maximize the QoS under a set of real-time and energy supply constraints (e.g., Aydin et al. 2001; Cortes et al. 2006; Yu et al. 2008, 2013a, b; Tchamgoue et al. 2013; Isabel et al. 2017; Zhou et al. 2017; Wei et al. 2017; Mo et al. 2017, 2018b). Specifically, QoS-aware task mapping determines on which processor should the task be executed (task-to-processor allocation), its frequency (frequency-to-task assignment), the start time (task scheduling), and the end time (optional subtask adjustment) of each task, such that the system QoS is maximized, while meeting the energy supply and the task deadlines. Usually, the task-to-processor allocation and frequency-to-task assignment decisions are *binary* variables, while the task scheduling and optional subtask adjustment

decisions can be either *integer* (Yu et al. 2008, 2013a, b; Isabel et al. 2017; Zhou et al. 2017; Wei et al. 2017) or *continuous* (Aydin et al. 2001; Cortes et al. 2006; Tchamgoue et al. 2013; Mo et al. 2017, 2018b) variables. Based on different constraints and variables introduced into the problem formulation, the variables may be coupled with each other nonlinearly, i.e., MINLP. However, it is not desirable since solving MINLP is much more complex than solving MILP. In order to linearize the nonlinear items, the common methods include:

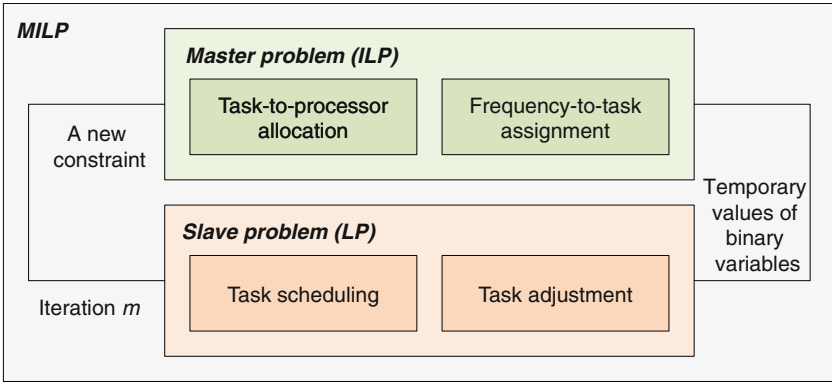
1. Linear approximation (Leung et al. 2003).
2. Variables replacement (Chen et al. 2014; Mo et al. 2018b).

The QoS-aware task mapping problem is similar to *Quadratic Assignment Problem*, a well-known NP-hard problem. Therefore, finding an optimal solution satisfying all the given constraints (e.g., energy efficiency, deadline, QoS, task dependency, and DVFS) is very difficult and time consuming (Mo et al. 2018b). Existing works usually focus on different contexts, as summarized in Table 1. The platforms considered in (Cortes et al. 2006; Yu et al. 2008; Tchamgoue et al. 2013) are single-core platforms, i.e., the task allocation problem is not taken into account. Although the works in (Aydin et al. 2001; Yu et al. 2013a, b; Isabel et al. 2017; Zhou et al. 2017; Wei et al. 2017; Mo et al. 2017) target at multi-core platforms, some assumptions are made during the problem formulation, e.g., the task-to-processor allocation is fixed and given in advance for all the tasks in (Yu et al. 2013a, b), the tasks considered in (Isabel et al. 2017) are independent, and in (Aydin et al. 2001; Zhou et al. 2017; Wei et al. 2017; Mo et al. 2017) each processor has a predefined frequency. The methods used to solve the aforementioned problems can be classified into two main classes:

1. The first class includes the methods based on heuristics (e.g., Yu et al. 2008, 2013a, b; Isabel et al. 2017; Zhou et al. 2017; Wei et al. 2017).

Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks, Table 1 Task mapping method

		Energy-aware						QoS-aware										
		(3)	(8)	(10)	(11)	(14)	(22)	(2)	(4)	(7)	(23)	(16)	(21)	(24)	(25)	(26)	(20)	(17)
Variables	Frequency-to-task	✓	✓	✓		✓	✓		✓	✓	✓			✓	✓			✓
	Task-to-processor	✓		✓	✓	✓		✓		✓		✓	✓			✓		✓
	Task adjustment							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Objective	Max. QoS							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Min. Energy	✓	✓	✓	✓	✓	✓			✓								
Constraints	Real-time	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Energy								✓		✓	✓	✓	✓	✓	✓		✓
Solution	Optimal	✓	✓					✓	✓			✓				✓	✓	
	Non-optimal			✓	✓	✓	✓			✓	✓		✓	✓	✓	✓		✓



Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks, Fig. 1 The structure of Benders decomposition

2. The second class includes the methods that always produce an optimal solution (e.g., Aydin et al. 2001; Cortes et al. 2006; Tchamgoue et al. 2013; Mo et al. 2017, 2018b).

Although heuristic methods can find feasible solutions in a short amount of time, they do not provide the bounds on the solution quality. When new assumptions or constraints are taken into account they must be redeveloped, and, thus, they are not always easily extensible.

The heuristic methods mentioned above usually adopted a multi-step optimization, i.e., to decouple the variables and to determine their values in sequence. For instance, a two-step heuristic called Adaptive Task Allocation (ATA) is pro-

posed in (Zhou et al. 2017). The aim of the first step is to find a task-to-processor allocation, such that the energy consumption is minimized. With the given task allocation decision, the energy consumed to execute one task is proportional to the length of its optional subtask. Based on the given task allocation decision, the aim of the second step is to adjust the optional subtasks so as to maximize OoS under the energy supply constraint. On the other hand, in order to find the optimal solution, the common methods include:

1. Convex optimization (Aydin et al. 2001; Cortes et al. 2006).
2. Benders Decomposition (BD) (Mo et al. 2017, 2018b).

The structure of BD is shown in Fig. 1. Instead of solving the binary and the continuous variables of MILP problem simultaneously, BD technique decomposes the original problem into two smaller problems with less variables and constraints: an ILP-based *Master Problem* (MP) and a LP-based *Slave Problem* (SP), and then solves them by utilizing the solution of one in the other. By doing so, the computation time can be significantly reduced. In each iteration, the current MP is solved to determine a lower bound for the original problem along with the temporary values of the binary variables. And then the SP is solved to obtain an upper bound by utilizing the MP solution. The bounds are updated if the stopping criterion, i.e., the gap between the upper and lower bounds is smaller than a predefined threshold, is not met, and a new constraint (Benders cut) is generated using the SP solution and is added to MP in next iteration.

Cross-References

- [Data Gathering in Wireless Sensor Networks](#)
- [Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks](#)
- [Middleware for Wireless Sensor Networks](#)

References

- Aydin H, Melhem R, Mosse D (2000) Tolerating faults while maximizing reward. In: IEEE Euromicro conference on real-time systems (EMRTS), IEEE pp 219–226
- Aydin H, Melhem R, Mosse D, Mejia-Alvarez P (2001) Optimal reward-based scheduling for periodic real-time tasks. *IEEE Trans Comput* 50(2):111–130
- Chen G, Huang K, Knoll A (2014) Energy optimization for real-time multiprocessor system-on-chip with optimal DVFS and DPM combination. *ACM Trans Embed Comput Syst* 13(3):1–21
- Cortes LA, Eles P, Peng Z (2006) Quasi-static assignment of voltages and optional cycles in imprecise computation systems with energy considerations. *IEEE Trans Very Large Scale Integr Syst* 14(10):1117–1129
- Dhasian HR, Balasubramanian P (2013) Survey of data aggregation techniques using soft computing in wireless sensor networks. *IET Inf Secur* 7(4): 336–342
- Esch J (2013) A survey on topology control in wireless sensor networks: taxonomy, comparative study, and open issues. *Proc IEEE* 101(12): 2534–2537
- Isabel M, Javier O, Rodrigo S, Paula Z (2017) Energy-aware scheduling mandatory/optional tasks in multi-core real-time systems. *Int Trans Oper Res* 24(12): 173–198
- Kong F, Yi W, Deng Q (2011) Energy-efficient scheduling of real-time tasks on cluster-based multicores. In: ACM/IEEE design, automation test in Europe (DATE), IEEE pp 1–6
- Lai X, Ji X, Zhou X, Chen L (2018) Energy efficient link-delay aware routing in wireless sensor networks. *IEEE Sensors J* 18(2):837–848
- Leung LF, Tsui CY, Ki WH (2003) Simultaneous task allocation, scheduling and voltage assignment for multiple-processors-core systems using mixed integer nonlinear programming. In: IEEE international symposium on circuits and systems (ISCAS), IEEE pp 309–312
- Li D, Wu J (2015) Minimizing energy consumption for frame-based tasks on heterogeneous multiprocessor platforms. *IEEE Trans Parallel Distrib Syst* 26(3):810–823
- Liu JWS, Shih WK, Lin KJ, Bettati R, Chung JY (1994) Imprecise computations. *Proc IEEE* 82(1): 83–94
- Liu X, Zhou H, Xiang J, Xiong S, Hou KM, de Vaulx C, Wang H, Shen T, Wang Q (2018) Energy and delay optimization of heterogeneous multicore wireless multimedia sensor nodes by adaptive genetic-simulated annealing algorithm. *Wirel Commun Mob Comput* 2018:1–13
- Mahmood A, Khan SA, Albalooshi F, Awwad N (2017) Energy-aware real-time task scheduling in multiprocessor systems using a hybrid genetic algorithm. *Electronics* 6(2):40
- Mo L, Kritikakou A, Sentieys O (2017) Decomposed task mapping to maximize QoS in energyconstrained real-time multicores. In: IEEE international conference on computer design (ICCD), IEEE pp 493–500
- Mo L, Cao X, Song Y, Kritikakou A (2018a) Distributed node coordination for real-time energyconstrained control in wireless sensor and actuator networks. *IEEE Internet Things, IEEE J*:1–12. <https://doi.org/10.1109/JIOT.2018.2839030>
- Mo L, Kritikakou A, Sentieys O (2018b) Controllable QoS for imprecise computation tasks on DVFS multicores with time and energy constraints. *IEEE J Emerg Sel Topics Circuits, IEEE Syst*:1–14. <https://doi.org/10.1109/JETCAS.2018.2852005>
- Munir A, Gordon-Ross A, Ranka S (2014) Multi-core embedded wireless sensor networks: architecture

- and applications. *IEEE Trans Parallel Distrib Syst* 25(6):1553–1562
- Silva R, Silva JS, Boavida F (2014) Mobility in wireless sensor networks: a survey and proposal. *Comput Commun* 52:1–20
- Tchamgoue GM, Kim KH, Jun YK, Lee WY (2013) Compositional real-time scheduling framework for periodic reward-based task model. *J Syst Softw* 86(6):1712–1724
- Wei T, Zhou J, Cao K, Cong P et al (2017) Cost-constrained QoS optimization for approximate computation real-time tasks in heterogeneous MPSoCs. *IEEE Trans Comput-Aided Des Integr Circuits Syst* 1(1):1–14
- Xu H, Kong F, Deng Q (2012) Energy minimizing for parallel real-time tasks based on level-packing. In: *IEEE international conference on embedded and real-time computing systems and applications (RTCSA)*, IEEE pp 98–103
- Yu H, Veeravalli B, Ha Y (2008) Dynamic scheduling of imprecise-computation tasks in maximizing QoS under energy constraints for embedded systems. In: *IEEE Asia and South Pacific design automation conference (ASP-DAC)*, IEEE pp 452–455
- Yu H, Ha Y, Veeravalli B (2013a) Quality-driven dynamic scheduling for real-time adaptive applications on multiprocessor systems. *IEEE Trans Comput*, IEEE 62(10):2026–2040
- Yu H, Veeravalli B, Ha Y, Luo S (2013b) Dynamic scheduling of imprecise-computation tasks on real-time embedded multiprocessors. In: *IEEE international conference on computational science and engineering (CSE)*, IEEE pp 770–777
- Zhou J, Yan J, Wei T, Chen M, Hu XS (2017) Energy-adaptive scheduling of imprecise computation tasks for QoS optimization in real-time MPSoC systems. In: *ACM/IEEE design, automation test in Europe (DATE)*, IEEE pp 1402–1407

Imprecise Computations

- [Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks](#)

Incentive Compatibility

- [Budget Feasible Mechanisms](#)

Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement

Xiaoyan Wang¹, Masahiro Umehira¹,
Biao Han², and Yusheng Ji³

¹Ibaraki University, Hitachi, Japan

²National University of Defense Technology,
Changsha, China

³Information Systems Architecture Science
Research Division, National Institute of
Informatics, Tokyo, Japan

Synonyms

[Auction mechanism](#); [Sensing augmented](#);
[Spectrum database](#)

Definitions

Spectrum measurement is of great importance to realize the dynamic spectrum sharing framework. To construct the spectrum database, spectrum measurement is needed which measures the power of the spectrum of desired signals. Generally, this operation is performed by dedicated equipments owned by the operators. Crowdsourcing the spectrum measurement tasks to the mobile users is a promising solution to improve the accuracy of the spectrum database in a cost-efficient way.

Background

With the tremendous increase of wireless devices and applications, there is a common belief that we are facing a severe shortage of spectrum resource for wireless communications in the near future. Currently, the spectrum resources below 6 GHz is almost fully allocated to primary users (PUs) in a static and exclusive way, which is highly

inefficient. To solve the spectrum crunch, we need a paradigm shift from the static, exclusive-use framework toward a dynamic spectrum sharing framework between PUs and secondary users (SUs). One promising solution that is being widely investigated is to employ a spectrum occupancy database, which relies on propagation models to calculate the received signal strength (RSS) at any receiver location. Unfortunately, this model-based spectrum database is prone to offer inaccurate and stale spectrum availability in many circumstances, e.g., approximately 40–70% of available white space is wasted in New York City (Saeed et al. 2014).

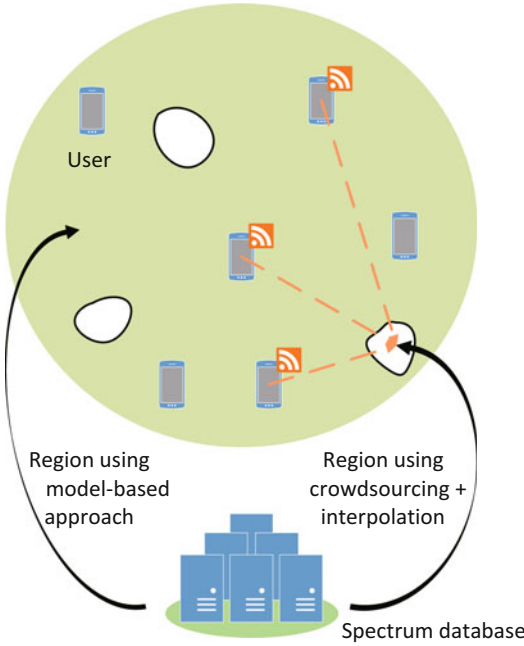
To improve the spectrum database accuracy, real-time spectrum measurements could be incorporated into the quasi-static database. A basic approach is to deploy dedicated sensors uniformly over the region of interest (Phillips et al. 2012). However it suffers from the high cost for database operator. A practical alternative is to exploit crowdsourcing for the sensing task (Nika et al. 2014), i.e., recruiting users with mobile devices that are outfitted with spectrum sensors. However participating in a crowdsourcing task may incur additional bandwidth usage and energy consumption for mobile users, and thus rewards in the form of either money or resource are needed to encourage them to make contributions.

Incentive mechanism for crowdsourcing-based spectrum measurement has been firstly investigated by Ying et al. (2015). Their goal is to minimize the interpolation variance for all the spots of interest for a given budget. Both budget-free mechanism with a cardinality constraint and budget-feasible mechanism by using bisection method are proposed. Gao et al. (2016) proposed a game-theoretic model-based mechanism to incentivize the users with additional spectrum access opportunities. Specifically, a two-level game model is considered, in which the database conducts dynamic pricing in a first-level Stackelberg game and SUs strategically contribute to spectrum sensing in a second-level stochastic game. One limitation of these mechanisms (Ying et al. 2015; Gao et al. 2016) is that they do not take into consideration the heterogeneity of spots that needs to be augmented and thus cannot recruit users to meet

distinct interpolation requirements. In Wang et al. (2017b) proposed a fine-grained incentive mechanism for sensing augmented spectrum database by utilizing auction model Wang et al. (2017a), where users sell their location-specific spectrum measurement to the database operator. Instead of trying to augment the whole spectrum database, the spectrum sensing is only conducted on spots with poor propagation model-based estimation. In Wang et al. (2017c) proposed a barter-like exchange model to incentivize the SUs by using spectrum access right. Based on this model, a truthful reverse auction approach is adopted to select the SUs and determine the individual access time in a computational efficient way. In Chen et al. (2017) proposed a reputation-based cooperative spectrum sensing incentive framework, where the cooperation stimulation problem is modeled as an indirect reciprocity game. In the proposed framework, SUs choose how to report their sensing results to the database center and gain reputations, based on which they can access a certain amount of vacant licensed channels in the future.

Crowdsourcing Augmented Spectrum Database

In the current propagation model-based spectrum database, the location-specific RSS is calculated based on the transceiver distance information. However, the accuracy of the database is low especially in the urban areas due to the influence of buildings. For the approaches that try to completely replace the propagation model-based database by using extensive real-time sensing results, its capital expenditure and operating cost are extremely high. To this end, crowdsourcing augmented spectrum database system is proposed (architecture is illustrated in Fig. 1), which consists of a centralized spectrum database and multiple distributed users with mobile devices that are outfitted with spectrum sensors. The calibration on RSS by using real-time sensing results could only be executed in spots with poor calculation accuracy (white spots in Fig. 1), and the rest of the areas still use the propagation model-based approach. By this method, the spectrum



Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement, Fig. 1 Architecture of the crowdsourcing augmented spectrum database

database accuracy could be improved with only a modest sensing effort. The spots that need crowdsourcing augmented could be separated out by using machine learning classifier, e.g., decision tree classifier in Chakraborty and Das (2014). Specifically, by exploring the features of distance from transmitter, received power, and line-of-sight (LOS) obstruction length, it is possible to learn where the database is likely to make inaccurate RSS calculations.

However, it is unlikely that a crowdsourcing participant locates exactly at the spot that the system wants to augment. Generally, the RSS for a desired spot is interpolated by using sensing measurements that are collected from surrounding users. Kriging Cressie and Cassie (1993) shows superior performance among different interpolation techniques and thus is widely utilized. Kriging is a linear regression method which uses location-specific known values to predict the unknown value in desired location. Consider that there is a random field in two dimensions and the value of that field (e.g., RSS in this entry) at a point (x_i, y_i) is $z_i = z(x_i, y_i)$. Given the values at a set of

locations $\mathcal{N} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, Kriging could predict the unknown value at a new location ($\hat{z}_0$) from the weighted known values as follows:

$$\hat{z}_0 = \sum_{i=1}^n \lambda_i z_i, \quad (1)$$

where λ_i is the normalized weight, i.e., $\sum_{i=1}^n \lambda_i = 1$. The optimal weights λ_i are determined by minimizing the estimation variance, i.e.:

$$\min_{\lambda_i} \text{Var}(\hat{z}_0 - z_0). \quad (2)$$

For the widely used ordinary Kriging (OK), it assumes z_i is intrinsically stationary such as:

$$E[z_i] = c \quad (3)$$

$$\text{Var}[z_i] = \sigma^2.$$

Let $\phi_{(x_0, y_0)}(\mathcal{N})$ denote the goal function in Eq. (2), which is the estimation variance at (x_0, y_0) by using the measurements at set $\mathcal{N}$. By substituting Eq. (1) into it, the following formula could be obtained:

$$\begin{aligned} \phi_{(x_0, y_0)}(\mathcal{N}) &= \text{Var} \left(\sum_{i=1}^n \lambda_i z_i - z_0 \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C_{ij} - 2 \sum_{i=1}^n \lambda_i C_{i0} + C_{00}, \end{aligned} \quad (4)$$

where $C_{ij} = \text{Cov}(z_i, z_j)$ and $\text{Cov}()$ denotes the covariance function.

To find minimum $\phi_{(x_0, y_0)}(\mathcal{N})$, a key function, namely, *semivariogram* γ_{ij} , is introduced. γ_{ij} models the variance between two points as a function of their distance. The theoretical semivariogram is represented by

$$\gamma_{ij} = \frac{1}{2} E \left[(z_i - z_j)^2 \right]. \quad (5)$$

Based on the OK assumption in Eq. (3), γ_{ij} could be expressed as

$$\gamma_{ij} = \sigma^2 - C_{ij}. \quad (6)$$

Substituting Eq. (6) into Eq. (4), $\phi_{(x_0, y_0)}(\mathcal{N})$ could be expressed by using semivariogram γ_{ij} as

$$\phi_{(x_0, y_0)}(\mathcal{N}) = 2 \sum_{i=1}^n \lambda_i \gamma_{i0} - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma_{ij} - \gamma_{00}. \quad (7)$$

By using Lagrange multiplier method, the minimization problem in Eq. (7) leads to solve the following matrix equation:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \dots & \gamma_{nn} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix} \quad (8)$$

where μ is the Lagrange parameter. Obviously, the optimal weight λ_i could be obtained if semivariogram γ_{ij} is known. In practice, γ_{ij} is estimated from measurements and then fitted with empirical curve, e.g., exponential or spherical model. The minimized $\phi_{(x_0, y_0)}(\mathcal{N})$ represents the estimation uncertainty at an unmeasured location (x_0, y_0) by using measurements at set $\mathcal{N}$, which could be used as a criterion in incentive mechanism design.

Incentive Mechanism

In this subsection, we introduce two kinds of incentive mechanisms for crowdsourcing-based spectrum measurement by using monetary reward and barter-like resource exchange, respectively. The considered network consists of a set of mobile users $\mathcal{N}$ with user index i , who knows its current location (x_i, y_i) . The set of spots that needs to be augmented (i.e., interpolated) by sensing results is denoted by $\mathcal{M}$ with spot index j . The result of interpolation is quantified by the Kriging estimation variance, e.g., the estimation variance at spot j by the interpolation of the user set $\mathcal{N}$ is denoted as $\phi_j(\mathcal{N})$. The spectrum database acquires the sensing data from users periodically. The set of spots that needs to be interpolated varies at each period. At the beginning of a period, the spectrum database

announces a sensing request with the center frequency information. Notice that the request does not need to contain the location information of the spots. Upon receiving the request, interested user i competes for this crowdsourcing task by submitting its location information (x_i, y_i) . The database selects a winner set $\mathcal{W}$ ($\mathcal{W} \subseteq \mathcal{N}$) to perform the crowdsourcing, and accordingly the informed winners report their sensing measurements to the database.

Monetary Reward-Based Incentive Mechanism

While participating in crowdsourcing, there is a cost c_i occurring to user i , which is related to the additional bandwidth usage and energy consumption. Notice that the cost is a private information and thus is only known by the user itself. If user could receive a monetary reward p_i from the database which is higher than its cost c_i , user i would have the incentive to participate in the crowdsourcing task. Therefore, based on the assumption that the users are rational, they are always trying to maximize their utility which could be defined as follows:

$$u_i = \begin{cases} p_i - c_i, & i \in \mathcal{W} \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

In a fine-grained system, each augmented spot j may have an estimation quality requirement r_j . r_j could be quantified by the Kriging estimation variance, that is, spot j 's RSS needs to be interpolated with a maximum estimation variance r_j . Therefore, the considered problem could be formulated as the following optimization problem:

$$\begin{aligned} & \min_{i \in \mathcal{W}} c_i \\ & \text{subject to } \phi_j(\mathcal{W}) \leq r_j, \forall j \in \mathcal{M}. \end{aligned} \quad (10)$$

where the objective function is to minimize the total costs of the selected user set $\mathcal{W}$ for crowdsourcing and the constraint function ensures that the quality requirement for each spot is met, i.e., the estimation variance at spot j by the interpolation of the winner set $\phi_j(\mathcal{W})$

is no larger than j 's requirement. The considered minimization problem is equivalent to the subset selection problem, which is proven to be NP-hard. Therefore, it is impossible to compute the optimal set of selected users that minimizes the total costs in polynomial time.

To solve this problem in an efficient way, an auction-based mechanism could be used (Wang et al. 2017b). Specifically, the database acts as a buyer who wants to buy measurement data, and the users act as sellers. Along with the location information, interested users also send bid b_i to the database, which may or may not be its real cost c_i . The mechanism consists of two phases:

winner selection and *payment determination*. Winner selection algorithm is based on the greedy heuristics that keeps selecting the next user with most *cost-efficient contribution* until all the spots' requirements are met. The cost-efficient contribution of user i changes according to the current winner set $\mathcal{W}$, which is defined as

$$\alpha_i(\mathcal{W}) = \frac{m_i(\mathcal{W})}{b_i}, \quad (11)$$

where b_i is i 's bid and $m_i(\mathcal{W})$ is i 's *total weighted marginal contribution* based on the current $\mathcal{W}$. $m_i(\mathcal{W})$ is calculated by

$$m_i(\mathcal{W}) = \sum_{j \in \mathcal{M}} \frac{\max(\phi_j(\mathcal{W}), r_j) - \max(\phi_j(\mathcal{W} \cup \{i\}), r_j)}{r_j}, \quad (12)$$

Here, the numerator $\max(\phi_j(\mathcal{W}), r_j) - \max(\phi_j(\mathcal{W} \cup \{i\}), r_j)$ could be regarded as the "marginal contribution" of user i to spot j under the current winner set $\mathcal{W}$. And this marginal contribution is weighted by $1/r_j$, i.e., the spot with small variance requirement has large weight.

To determine the payment, the key point is to find the maximum bid each winner can submit that allows it to win, i.e., critical payment. By using the critical payment, the auction mechanism could be proven to be truthful. Truthful is one of the most important properties that an auction mechanism desires, since in a truthful auction, bidding the true cost ($b_i = c_i$) is the dominant strategy regardless of other users' strategies. Specifically, to find the critical payment for winner k , the payment determination algorithm repeats the winner selection algorithm for the user set without k , i.e., $\mathcal{N}' = \mathcal{N} \setminus \{k\}$. In set $\mathcal{N}'$, the mechanism finds the winner l' who has the maximum $\alpha_{l'}(\mathcal{W}')$ based on the current winner set $\mathcal{W}'$. To let winner k replace winner l' , its cost-efficient contribution needs to be larger than that of l' , that is:

$$\frac{m_k(\mathcal{W}')}{b_k} > \frac{m_{l'}(\mathcal{W}')}{b_{l'}}. \quad (13)$$

In other words, the maximum bid for winner k that lets it replace winner l' is $\frac{m_k(\mathcal{W}')}{m_{l'}(\mathcal{W}')} \cdot b_{l'}$. Eventually, the maximum of these values is used as the critical payment for k .

Barter-Like Resource Exchange-Based Incentive Mechanism

It is assumed that the database incentivize users by using a nonmonetary payment p_i which is in the form of additional channel access time. And each user that participates the crowdsourcing task has a desired channel access time, which is denoted by t_i . Notice that t_i is a private information and thus is only known by the user itself. The length of each round that the database operator wants to collect new measurements is T , during which the selected users could share the channel access in a TDMA fashion. The rational user always tries to maximize its utility, which is defined as

$$u_i = \begin{cases} p_i - t_i, & i \in \mathcal{W} \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

Therefore, the considered problem could be formulated as the following optimization problem:

$$\begin{aligned} & \min_{\mathcal{W}} \sum_{j \in \mathcal{M}} \phi_j(\mathcal{W}) \\ & \text{subject to} \quad \sum_{i \in \mathcal{W}} p_i \leq T, \end{aligned} \quad (15)$$

where the objective function is to minimize the total Kriging variance for spot set $\mathcal{M}$ by the interpolation of selected winner set $\mathcal{W}$, and the constraint ensures that the sum of the payment for winner set $\mathcal{W}$ does not exceed the time interval T . Similar to Eq. (1), this optimization problem is NP-hard, and an auction-based mechanism could solve it in an efficient way (Wang et al. 2017c).

The mechanism consists of three phases, i.e., winner selection, payment determination, and bidding threshold search. Winner selection algorithm is based on the greedy heuristics that keeps selecting the next user with highest *contribution-bidding ratio* until the total bids exceed the current *bidding threshold*. The bidding threshold is related to the time interval T , which could be found by employing a bisection searching method. The contribution-bidding ratio of user i based on the current winner set $\mathcal{W}$ is defined as

$$\alpha_i(\mathcal{W}) = \frac{c_i(\mathcal{W})}{b_i}, \quad (16)$$

where b_i is i 's bid and $c_i(\mathcal{W})$ is i 's *interpolation variance contribution* based on the current $\mathcal{W}$. $c_i(\mathcal{W})$ is calculated by

$$c_i(\mathcal{W}) = \sum_{j \in \mathcal{M}} (\phi_j(\mathcal{W}) - \phi_j(\mathcal{W} \cup \{i\})). \quad (17)$$

To find the critical payment for winner k , the winner selection algorithm for the user set without k is repeated, i.e., $\mathcal{N}' = \mathcal{N} \setminus \{k\}$. In set $\mathcal{N}'$, the winner k' could be found who has the maximum $\alpha_{k'}(\mathcal{W}')$ based on the current winner set $\mathcal{W}'$. To let winner k replace winner k' , its contribution-bidding ratio needs to be larger than that of k' , that is:

$$\alpha_k(\mathcal{W}') = \frac{c_k(\mathcal{W}')}{b_k} > \alpha_{k'}(\mathcal{W}') = \frac{c_{k'}(\mathcal{W}')}{b_{k'}}. \quad (18)$$

In other words, the maximum bid for winner k that lets it replace winner k' is $\frac{c_k(\mathcal{W}')}{c_{k'}(\mathcal{W}')} \cdot b_{k'}$. Eventually, after finishing the winner selection loop on user set $\mathcal{N}'$, the maximum of these values is used as the critical payment for k .

In a truthful auction, generally the final total payments are larger than the sum of winners' bids. The final total payment is constrained by the time interval T . However, it is impossible to use constraint T to control the winner selection process, since the payment is determined after the selection of winners. Therefore, it is required to define a bidding threshold H to control the winner selection loop, i.e., the sum of the bids for all winners should not exceed this threshold. Obviously, the bidding threshold is smaller than the time constraint T , and using a bisection searching method could find it. Specifically, a lower bound l and an upper bound u are used to narrow down the possible range of the bidding threshold, and a tolerance e is employed to stop the searching process.

Key Applications

The crowdsourcing-based spectrum measurements are crucial for the dynamic spectrum sharing. Due to the tremendous increasing of mobile data traffic, dynamic and efficient spectrum sharing mechanism is required. To realize interference protection and reuse the whitespace efficiently, an accurate spectrum database that using crowdsourced spectrum measurement is promising.

Cross-References

- [Auction](#)
- [Database-Based Spectrum Access](#)
- [Incentive Mechanisms for Mobile Data Crowdsourcing](#)

References

- Chakraborty A, Das SR (2014) Measurement-augmented spectrum databases for white space spectrum. In: Proceedings of the 10th ACM international on conference on emerging networking experiments and technologies. ACM, New York, CoNEXT '14, pp 67–74. <https://doi.org/10.1145/2674005.2675001>
- Chen B, Zhang B, Yu JL, Chen Y, Han Z (2017) An indirect reciprocity based incentive framework for cooperative spectrum sensing. In: 2017 IEEE international conference on communications (ICC), pp 1–6. <https://doi.org/10.1109/ICC.2017.7996606>
- Cressie NA, Cassie NA (1993) Statistics for spatial data. Wiley, New York
- Gao B, Bhattarai S, Park JMJ, Yang Y, Liu M, Zeng K, Dou Y (2016) Incentivizing spectrum sensing in database-driven dynamic spectrum sharing. In: 2016 IEEE conference on computer communications (INFOCOM), pp 1–9. <https://doi.org/10.1109/INFOCOM.2016.7524586>
- Nika A, Zhang Z, Zhou X, Zhao BY, Zheng H (2014) Towards commoditized real-time spectrum monitoring. In: Proceedings of the 1st ACM workshop on hot topics in wireless. ACM, New York, HotWireless '14, pp 25–30. <https://doi.org/10.1145/2643614.2643615>
- Phillips C, Ton M, Sicker D, Grunwald D (2012) Practical radio environment mapping with geostatistics. In: 2012 IEEE international symposium on dynamic spectrum access networks, pp 422–433. <https://doi.org/10.1109/DYSPAN.2012.6478166>
- Saeed A, Harras KA, Youssef M (2014) Towards a characterization of white spaces databases errors: an empirical study. In: Proceedings of the 9th ACM WiTECH '14. ACM, New York, pp 25–32. <https://doi.org/10.1145/2643230.2643234>
- Wang X, Ji Y, Zhou H, Li J (2017a) Auction-based frameworks for secure communications in static and dynamic cognitive radio networks. IEEE Trans Veh Technol 66(3):2658–2673. <https://doi.org/10.1109/TVT.2016.2581179>
- Wang X, Umehira M, Li P, Gu Y, Ji Y (2017b) Fine-grained incentive mechanism for sensing augmented spectrum database. In: GLOBECOM 2017 – 2017 IEEE global communications conference, pp 1–6. <https://doi.org/10.1109/GLOCOM.2017.8254447>
- Wang X, Umehira M, Li P, Gu Y, Ji Y (2017c) Incentivizing crowdsourcing for exclusion zone refinement in spectrum sharing system. In: 2017 23rd Asia-Pacific conference on communications (APCC), pp 1–6. <https://doi.org/10.23919/APCC.2017.8303975>
- Ying X, Roy S, Poovendran R (2015) Incentivizing crowdsourcing for radio environment mapping with statistical interpolation. In: 2015 IEEE international symposium on dynamic spectrum access networks (DySPAN), pp 365–374. <https://doi.org/10.1109/DySPAN.2015.7343932>

Incentive Mechanisms for Mobile Crowdsensing

Xiaowen Gong

Auburn University, Auburn, AL, USA

Synonyms

Incentive mechanisms for mobile data crowdsourcing

Definition

Incentive mechanisms incentivize mobile users to behave as desired by the requester of mobile crowdsensing

Background

Mobile crowdsensing has found a wide range of applications, such as spectrum sensing (Ding et al. 2014) and environmental monitoring (e.g., air quality Mun et al. (2009), noise level Rana et al. (2010), and weather conditions Atmos (2014) like temperature, humidity, and wind speed). In principle, crowdsensing (For brevity, we call “mobile crowdsensing” as “crowdsensing”) leverages the “wisdom” of a potentially large crowd of mobile users by allocating a sensing task to them and collecting the data from them for the task. The popularity of crowdsensing is promoted by powerful mobile devices equipped with various sensors and pervasive Internet connectivity. This trend is expected to grow in the foreseeable future with the emerging Internet of Things (IoT). With enormous opportunities brought by big data, crowdsensing serves as an important first step for data mining tools to harness the power of big data in many application domains.

Although the benefit of crowdsensing is pronounced, performing a sensing task typically incurs overhead for the participating user, in terms of the user’s resource consumption devoted

to sensing, such as battery and computing power. Further, the participating user also incurs the risk of potential privacy loss by sharing its sensed data with others. In general, a user may not participate in sensing without receiving adequate incentive. Therefore, effective incentive design is essential for realizing the benefit of crowdsensing.

Overview

The past few years have seen many studies on incentive mechanisms for crowdsensing (see, e.g., Duan et al. 2012; Shah and Zhou 2015; Wang et al. 2016). Most of these works use monetary reward to stimulate users' participation, where the crowdsensing requester pays some money to a user for participating in crowdsensing. A well-known real-world example is Amazon Mechanical Turk (MTurk), where a user receives payment (e.g., \$0.2) for completing a crowdsensing task. Another class of incentive mechanisms (e.g., see Luo and Tham 2012; Gong et al. 2015) exploits users' service to stimulate users' participation. In this situation, a user acts as both a contributor and a consumer, such that the service it receives as the consumer compensates the service it provides as the contributor. A typical real example is traffic monitoring (e.g., Google Maps, Waze), where a user provides its location which is used to estimate real-time traffic information, while the user itself also receives this information estimated from the locations provided by other users. Social relationships have also been exploited Gong et al. (2015) to incentivize users to perform crowdsensing tasks for their social friends, as users are typically altruistic to their social friends. In this article, we focus on incentive mechanisms based on monetary reward.

A key challenge in the incentive design for crowdsensing is that a user may have private information (e.g., cost, data, quality) or hidden actions (e.g., effort) which is unknown to and cannot be verified by the crowdsourcing requester. As a result, a strategic user may have incentive to manipulate its private information revealed to the requester or its hidden actions,

so as to gain an advantage. In the following, we will discuss truthful mechanisms that incentivize users to truthfully report their private information to the requester and take hidden actions as desired by the requester.

Classification of Truthful Mechanisms

Truthful Mechanisms for Cost Elicitation

There have been many mechanisms to incentivize users to truthfully reveal their cost of participation in crowdsensing (e.g., see Yang et al. 2012; Koutsopoulos 2013; Tarable et al. 2015; Luo et al. 2015). The cost is considered to be a strategic user's private information that it would not be willing to reveal truthfully to the user's advantage without appropriate incentive. Intuitively, a user would overreport its cost to the requester, in the hope of receiving a higher reward to compensate its cost. Most incentive design problems in this category are formulated as a reverse auction, where the requester is the auctioneer who pays users for performing crowdsensing tasks. Some studies (Zhang et al. 2016; Jin et al. 2017a) have considered a double auction where there are multiple requesters with multiple tasks for users to perform. Some other studies (Subramanian et al. 2015; Han et al. 2016, 2017) have considered an online auction where users arrive and leave sequentially in time. As a major characteristic of mobile crowdsensing, the locations and mobility of users have been taken into account explicitly in some works (Feng et al. 2014; Zhang et al. 2016). Many existing mechanisms for cost elicitation have been designed based on the classical VCG auction and the characterization of truthful mechanisms (Nisan et al. 2007, Theorem 9.36).

Truthful Mechanisms for Effort Elicitation

Incentive mechanisms for hidden actions have been well studied in the economics literature (Bolton and Dewatripont 2005). These studies are often concerned with strategic agents that can make hidden effort that are not desired by a principal who employs the agents to work on a task. Intuitively, a user would make less effort in the task so as to reduce its cost incurred

in the task. There are a few recent studies that have investigated this problem in the context of data crowdsourcing (Cai et al. 2015; Luo et al. 2015; Jin et al. 2017b). Cai et al. (2015) have proposed truthful mechanisms to incentivize users to make efforts as desired in statistical estimation. However, most of the work on truthful elicitation of effort do not take into consideration agents' other possible private information (e.g., quality, cost). Luo et al. (2015) have designed mechanisms that not only elicit desired efforts from users but also truthful revelation of their private cost and data. Liu and Chen (2016b, 2017) has designed effort elicitation mechanisms for workers with diverse cost and studied the optimal reward for maximizing the requester's payoff. Liu and Chen (2016a) has designed a bandit-based mechanism for sequential effort elicitation.

Truthful Mechanisms for Data Elicitation

Mechanism design for truthful elicitation of strategic agents' data (e.g., opinions) has been extensively studied in various applications (Prelec 2004; Miller et al. 2005), more recently in the context of data crowdsourcing (Dasgupta and Ghosh 2013; Luo et al. 2015). The data of an agent can be its private information that it can manipulate in favor of its benefit. To protect users' privacy, one effective method is to add noise to the data reported by users, based on some privacy metric (e.g., differential privacy). Some works (Wang et al. 2016) have studied mechanisms that incentivize users to report data of a privacy level desired by the requester.

Truthful Mechanisms for Quality Elicitation

Data accuracy is a key performance metric of crowdsourcing. This is determined by the quality of the individual user's data contributed to the task, which captures the accuracy of the user's data compared to the ground truth. In general, the quality of users' data varies for different users depending on a user's specific situation (e.g., location, environment). Intuitively, a user would overreport its quality to the requester, in the hope of receiving a higher reward for contributing

high-quality data. To fully exploit the potential of crowdsourcing, it is important to perform task allocation and data aggregation based on users' data quality. Incentive mechanisms for quality-based crowdsensing have been studied in a few works (Koutsopoulos 2013; Shah and Zhou 2015; Liu and Liu 2015; Jin et al. 2015, 2016, 2017a). One interesting line of work (Jin et al. 2015, 2016, 2017a) in this direction has studied truthful mechanisms for quality-based task allocation where workers have private participating cost. Recent works (Gong and Shroff 2017, 2018) have proposed a quality-aware crowdsourcing framework and devised truthful mechanisms that incentivize users to truthfully report their quality and data to the requester and make effort desired by the requester.

Cross-References

- Budget Feasible Mechanisms
- Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement
- Private Truth Discovery in Cloud-Empowered Crowdsensing Systems
- Social IoT Crowd-Sourcing on Disaster Reduction

References

- Atmos (2014) Atmos: crowd sensing weather platform. <http://beja.m-iti.org/web/?q=node/1>
- Bolton P, Dewatripont M (2005) Contract theory. MIT Press, Cambridge
- Cai Y, Daskalakis C, Papadimitriou CH (2015) Optimum statistical estimation with strategic data sources. In: Conference on learning theory (COLT)
- Dasgupta A, Ghosh A (2013) Crowdsourced judgement elicitation with endogenous proficiency. In: International World Wide Web conference (WWW)
- Ding G, Wang J, Wu Q, Zhang L, Zou Y, Yao YD, Chen Y (2014) Robust spectrum sensing with crowd sensors. *IEEE Trans Commun* 62(9):3129–3143
- Duan L, Kubo T, Sugiyama K, Huang J, Hasegawa T, Walrand J (2012) Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing. In: IEEE international conference on computer communications, Orlando, FL (INFOCOM)
- Feng Z, Zhu Y, Zhang Q, Ni LM, Vasilakos AV (2014) Trac: truthful auction for location-aware collabora-

- tive sensing in mobile crowdsourcing. In: IEEE international conference on computer communications, Toronto (INFOCOM)
- Gong X, Shroff N (2017) Truthful mobile crowdsensing for strategic users with private qualities. In: International symposium on modeling and optimization in mobile, ad hoc, and wireless networks, Paris (WiOpt)
- Gong X, Shroff N (2018) Incentivizing truthful data quality for quality-aware mobile data crowdsourcing. In: ACM international symposium on mobile ad hoc networking and computing, Los Angeles, CA (MobiHoc)
- Gong X, Chen X, Zhang J, Poor HV (2015) Exploiting social trust assisted reciprocity (star) toward utility-optimal socially-aware crowdsensing. *IEEE Trans Sig Inf Proces Over Netw* 1(3):195–208
- Han K, Zhang C, Luo J, Hu M, Veeravalli B (2016) Truthful scheduling mechanisms for powering mobile crowdsensing. *IEEE Trans Comput* 65(1):294–307
- Han K, He Y, Tan H, Tang S, Huang H, Luo J (2017) Online pricing for mobile crowdsourcing with multi-minded users. In: Proceedings of the 18th ACM international symposium on mobile ad hoc networking and computing. ACM, p 18
- Jin H, Su L, Chen D, Nahrstedt K, Xu J (2015) Quality of information aware incentive mechanisms for mobile crowd sensing systems. In: ACM international symposium on mobile ad hoc networking and computing, Hangzhou (MobiHoc)
- Jin H, Su L, Xiao H, Nahrstedt K (2016) INCEPTION: incentivizing privacy-preserving data aggregation for mobile crowd sensing systems. In: ACM international symposium on mobile ad hoc networking and computing, Paderborn (MobiHoc)
- Jin H, Su L, Nahrstedt K (2017a) CENTURION: incentivizing multi-requester mobile crowd sensing. In: IEEE international conference on computer communications, Atlanta, GA (INFOCOM)
- Jin H, Su L, Nahrstedt K (2017b) Theseus: incentivizing truth discovery in mobile crowd sensing systems. In: ACM international symposium on mobile ad hoc networking and computing, Chennai (MobiHoc)
- Koutsopoulos I (2013) Optimal incentive-driven design of participatory sensing systems. In: IEEE international conference on computer communications, Turin (INFOCOM)
- Liu Y, Chen Y (2016a) A bandit framework for strategic regression. In: Advances in neural information processing systems, pp 1821–1829
- Liu Y, Chen Y (2016b) Learning to incentivize: eliciting effort via output agreement. In: international joint conference on artificial intelligence (IJCAI)
- Liu Y, Chen Y (2017) Sequential peer prediction: learning to elicit effort using posted prices. In: AAAI conference on artificial intelligence (AAAI)
- Liu Y, Liu M (2015) An online learning approach to improving the quality of crowd-sourcing. In: ACM international conference on measurement and modeling of computer systems, Portland, OR (SIGMETRICS)
- Luo T, Tham CK (2012) Fairness and social welfare in incentivizing participatory sensing. In: 2012 9th annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks (SECON). IEEE, pp 425–433
- Luo Y, Shah NB, Huang J, Walrand J (2015) Parametric prediction from parametric agents. In: The 10th workshop on the economics of networks, systems and computation (NetEcon)
- Miller N, Resnick P, Zeckhauser R (2005) Eliciting informative feedback: the peer-prediction method. *Manag Sci* 51(9):1359–1373
- Mun M, Reddy S, Shilton K, Yau N, Burke J, Estrin D, Hansen M, Howard E, West R, Boda P (2009) PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In: ACM international conference on mobile systems, applications, and services (MobiSys)
- Nisan N, Roughgarden T, Tardos E, Vazirani VV (2007) Algorithmic game theory, vol 1. Cambridge University Press, Cambridge
- Prelec D (2004) A Bayesian truth serum for subjective data. *Science* 306(5695):462–466
- Rana RK, Chou CT, Kanhere SS, Bulusu N, Hu W (2010) Ear-phone: an end-to-end participatory urban noise mapping system. In: ACM/IEEE international conference on information processing in sensor networks (IPSN)
- Shah NB, Zhou D (2015) Double or nothing: multiplicative incentive mechanisms for crowdsourcing. In: Conference on neural information processing systems (NIPS)
- Subramanian A, Kanth GS, Moharir S, Vaze R (2015) Online incentive mechanism design for smartphone crowd-sourcing. In: International symposium on modeling and optimization in mobile, ad hoc, and wireless networks, Mumbai (WiOpt)
- Tarable A, Nordio A, Leonardi E, Marsan MA (2015) The importance of being earnest in crowdsourcing systems. In: IEEE international conference on computer communications, Hong Kong (INFOCOM)
- Wang W, Ying L, Zhang J (2016) The value of privacy: strategic data subjects, incentive mechanisms and fundamental limits. In: ACM international conference on measurement and modeling of computer systems, Antibes Juan-les-Pins (SIGMETRICS)
- Yang D, Xue G, Fang X, Tang J (2012) Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In: ACM annual international conference on mobile computing and networking, Istanbul (MobiCom)
- Zhang H, Liu B, Susanto H, Xue G, Sun T (2016) Incentive mechanism for proximity-based mobile crowd service systems. In: IEEE international conference on computer communications, San Francisco, CA (INFOCOM)

Incentive Mechanisms for Mobile Data Crowdsourcing

► [Incentive Mechanisms for Mobile Crowdsensing](#)

Index Modulation

Xiang Cheng¹, Miaowen Wen², and Liuqing Yang³

¹State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing, China

²School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

³Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA

Synonyms

[MIMO](#); [OFDM](#); [Spatial modulation](#)

Definition

Index modulation (IM) refers to a family of modulation techniques that rely on the activation states of some resources/building blocks for information embedding. The resources/building blocks can be either physical, e.g., antenna, sub-carrier, time slot, and frequency carrier, or virtual, e.g., virtual parallel channels, signal constellation, space-time matrix, and antenna activation order. A distinct feature of IM is that part of the information is implicitly embedded into the transmitted signal.

Historical Background

Wireless communications are the fastest growing segment of the communications industry. From around 1980 to the present, the wireless systems have traveled through 1st generation (1G), 2G, and 3G. Recently, the long-term evolution (LTE) and its extension, LTE-advanced (LTE-A), as practical 4G systems, are steadily expanding across global markets. Nowadays, 5G of wireless networks, which is expected to be introduced around 2020, has been one of the hottest topics in the wireless communications community over the past few years. Although there is an ongoing debate on 5G wireless technology whether it will be a simple evolution compared to its fourth generation (4G) counterpart or a radically new communication network, 5G wireless networks are expected to provide immense bandwidth and much higher data rates with considerably lower latency. However, the currently employed modulation techniques based on multi-input-multi-output orthogonal frequency division multiplexing (MIMO-OFDM), unfortunately, are not sufficient in satisfying such 5G requirements. Conventional MIMO may achieve high spectrum efficiency (SE) with massive antennas but has compromised energy efficiency (EE) due to the scaled power consumption of a large number of radio-frequency (RF) chains. The OFDM modulation is prone to Doppler-induced intercarrier interference (ICI). Its inherent high peak-to-average power ratio (PAPR) also necessitates expensive power amplifiers. Novel advanced modulation techniques are therefore very much needed. To this end, the recently emerging IM techniques arise as a promising candidate that has the potential to meet the 5G requirements.

Foundations

Existing IM schemes are mainly carried out in space, time, and frequency slots or the combination of them, among which the most marked similarity lies in the utilization of the active radio

resources to carry information. Generally speaking, IM splits the information bits into index bits and constellation bits, where the former determine which portion of the radio resources (antennas, subcarriers, etc.) is active and the latter are mapped to conventional constellation symbols that are to be carried by the active resources. Specifically, the index mapping of existing IM schemes in different domains is classified as follows:

- *IM in Space.* Spatial modulation (SM) is a representative IM technique in the space domain, which works with a single RF chain and conveys information via the antenna index (Cheng et al. 2018; Mesleh et al. 2008; Di Renzo et al. 2014). The SM signal constellation consists of (1) the conventional quadrature amplitude modulation constellation and (2) the antenna index constellation. Conveying information via the active antenna index, SM enjoys spatial multiplexing gain with a single-RF chain transmitter. Hence, SM is more energy efficient, and its detection complexity is relatively lower. Note that one drawback of SM is that the information conveyed by the antenna index increases in $\log_2 N_t$, where N_t is the number of transmit antennas. When N_t is large, SM suffers from compromised SE. The idea of SM can also accommodate multiple active transmit antennas to improve SE, giving rise to generalized SM (G-SM) (Younis et al. 2010; Wang et al. 2012). However, it requires multiple RF chains and induces inter-channel interference.
- *IM in the Space-Time.* Transmitting signals across multiple time slots facilitate transmit diversity in MIMO systems. Previous focus was on space-time matrix designs that strike the largest coding and diversity gains with satisfactory receiver complexity. Recently, interests shift the exploration of the space-time resource to convey information. Differential SM (D-SM), an IM technique in the space-time domain, is a representative example (Bian et al. 2015). The index mapping of D-SM determines the current antenna activation order according to that in the former space-time block and the current index bits. As a differential solution to SM, D-SM avoids the challenging channel estimation inherent to SM. Like SM, D-SM operates with a single RF chain but bypasses the accurate estimation of the channel state information (CSI) that is computationally demanding.
- *IM in Frequency.* OFDM-IM is a representative IM technique in the frequency domain, which extends the SM principle to OFDM subcarriers (Frenger et al. 1999; Abualhiga et al. 2009; Basar et al. 2013). In OFDM-IM, not all subcarriers carry information symbols, and the indices of inactive subcarriers convey information via IM. To date, several works have demonstrated that OFDM-IM can outperform plain OFDM in terms of the bit error rate (BER) under the same SE. It is revealed by Basar et al. (2013) that (1) OFDM-IM achieves the maximum rate if and only if the subcarriers within each group experience independent fading via, e.g., interleaved grouping; (2) OFDM-IM with interleaved grouping can achieve an up to 3 dB signal-to-noise (SNR) gain over plain OFDM for small modulation cardinality M (typically 2, 4), though this improvement becomes smaller and even diminishes as M grows; and (3) the advantage of OFDM-IM over plain OFDM can be maximized by choosing a specific number of inactive subcarriers (typically 1, or 2).
- *IM in Space-Frequency.* MIMO-OFDM has already been widely accepted as one fundamental technique in current wireless communication systems. To incorporate IM into MIMO-OFDM, one straightforward way is to carry out conventional modulation in one domain (space or frequency/subcarrier) and superimpose IM in the other domain. For example, using all subcarriers while implementing independent SM per subcarrier leads to OFDM-SM, whereas using all antennas but implementing independent subcarrier activation per each antenna would result in MIMO-OFDM with IM (MIMO-OFDM-IM) (Basar et al. 2016).

Instead of treating space and frequency domains independently, generalized space-frequency IM (GSFIM) (Datta et al. 2016) is proposed to jointly select the active elements. In GSFIM, a group of transmit antennas are activated by the information bits, and the active space-frequency elements are then jointly selected according to the remaining information bits. Therefore, it is the generalization of OFDM-SM and MIMO-OFDM-IM.

Based on the fundamental concept of IM in the space, time, and frequency, various articles aim at the improvement of the SE/EE, the potential of transmit diversity, precoding, and other enhancements of current IM techniques. Due to the similarity among different IM techniques, the enhancements proposed for a specific scheme are usually applicable to all IM techniques. To better illustrate these, three typical enhancements designed for SM, which can be readily generalized to other IM schemes, are introduced in the following.

In-Phase/Quadrature IM

The RF signals usually contain the in-phase component and the quadrature component. The orthogonality between them can be exploited to improve the SE of IM techniques. For example, given a symbol drawn from the PSK/QAM constellation, one may map its real part with a cosine carrier into a randomly selected antenna and its imaginary part with a sine carrier to another randomly selected antenna. This technique, termed as quadrature SM (QSM) (Mesleh et al. 2015), retains the requirement of a single RF chain while allowing more spatial bits to be transmitted. The frequency domain counterpart of OFDM-IM in-phase/quadrature (OFDM-IM-IQ) is shown to have significantly better error performance than IM-OFDM at the same SE (Fan et al. 2015).

Precoded IM

IM is not limited to a transmitter design. With transmitter-side CSI, effective precoding techniques, such as zero-forcing and MMSE

precoders, can facilitate parallel interference-free channels at the receive antennas. Therefore, IM techniques can be transplanted to the receiver side. By selecting the active receive antenna indices, precoding-aided SM (PSM) has been proposed by Stavridis et al. (2016). Different from SM/G-SM, the detection of the constellation symbols at different antennas can be performed in a parallel manner, therefore significantly reducing the receiver complexity especially when multiple antennas are activated. Also, in order to bypass channel estimation in PSM, D-SM can also be applied at the receiver side, which also induces 3dB performance penalty. IM in the space-frequency domain can also be applied at the receiver via precoding. With precoding, the received signals at different receive antennas are free of interchannel interference. As a result, the joint selection of the active elements in space and frequency domains will not increase the detection complexity as at the transmitter side, which will allow more flexible grouping across the space and frequency domain to reduce the intragroup channel correlation and consequently improve the error performance.

Diversity-Enhancing IM

Beside the SE improvement and the receiver complexity reduction, research has been dedicated to the error performance enhancement of IM. Among these, diversity-enhancing IM schemes are attracting increasing interests. Since IM are naturally applicable to multiple resources (antennas/subcarriers), some well-known space-time block codes (STBC), such as Alamouti, can be integrated by carefully designing the mapping of index bits. For example, STBC-SM is proposed to exploit the transmit diversity by integrating STBC technique (Basar et al. 2011). Such designs are usually migratable among different IM schemes given similar channel properties. Specifically, the coordinate interleaving method proposed for IM-OFDM is also applicable to all precoded IM schemes, because, with precoding, the received signal among the receive antennas becomes interference-free.

Practical Issues for IM

Practical implementation issues for IM are demonstrated in possible application areas. More specifically, we focus our attention into underwater acoustic (UWA), vehicle to everything (V2X), and wireless-powered (WP) communications. The concept of activating partial subcarriers during the transmission of an OFDM symbol endows OFDM-IM and its variants with potential of adaptation to high mobility scenarios, such as V2X and UWA communications, since the power of the Doppler-induced ICI, which scales with the number of active subcarriers, will be significantly reduced.

The first attempt on applying OFDM-IM to V2X communications was made by Cheng et al. (2014), where great benefits of OFDM-IM with interleaved grouping in terms of BER and achievable rate performance are revealed via analytical and simulated comparisons with classical OFDM and OFDM-IM with localized grouping. Altalib et al. (2017) investigated the performance of OFDM-IM with all SAPs encoded in a V2X scenario and demonstrated that OFDM-IM can be considered as a strong candidate for the transmission technique of the vehicular communications standard.

The first application of OFDM-IM to UWA communications was reported by Wen et al. (2014). Noticing that the power leakage from the active subcarriers to the inactive ones because of ICI significantly increases the possibility of erroneous detection of subcarrier states, Wen et al. (2014) proposed a solution by integrating the well-known ICI self-cancellation techniques into the OFDM-IM framework. With a simple modification, OFDM-IM is shown to be very effective in ICI mitigation by both computer simulations and sea experiments. The integration idea proposed by Wen et al. (2014) is generalized by Wen et al. (2016), where both the one-path and two-path implementations are demonstrated. In addition, a hybrid OFDM-IM system with improved spectral efficiency, which introduces an additional bit to indicate the OFDM-IM mode or the classical OFDM mode, is proposed.

On the other hand, the SM concept is first applied at the wireless device into a WP communication system for information transmission (IT) by mapping information to the active antenna indices (Zhang and Cheng 2017). To be specific, this novel scheme enjoys the advantages of improved system achievable rate (AR), energy harvesting (EH) and information transmission (IT), and self-energy recycling

Index Modulation, Table 1 Comparison of existing IM schemes

Scheme	RF chain number	SE	TX/RX complexity	Diversity order	CSIT/CSIR	Key advantage
SM	1	Low	Low/Low	N_r	No/Yes	High EE
D-SM	1	Low	Low/Low	N_r	No/No	Bypass channel estimation and high SE
G-SM	$< N_t$	Medium	Medium/Medium	N_r	No/Yes	Flexibly SE
PSM	N_t	Low	High/Low	$N_t - N_r + 1$	Yes/Yes	Low Rx complexity
QSM	1	Medium	Low/Low	N_r	No/Yes	Moderate SE and EE
STBC-SM	2	Medium	Medium/Medium	$2N_r$	No/Yes	High reliability and moderate EE
OFDM-IM	1	Medium	Low/Low	N_r	No/Yes	High EE
OFDM-IM-IQ	1	Medium	Medium/Medium	N_r	No/Yes	High EE and moderate SE
MIMO-OFDM-IM	N_t	High	High/High	N_r	No/Yes	High EE and moderate SE
GSFIM	N_t	Medium	High/High	N_r	No/Yes	Flexible SE

N_t is the number of transmit antennas;

N_r is the number of receive antennas

(SE-R) while requiring no backhaul links between the WD and the information receiver.

Discussion

In order to comprehensively compare existing IM schemes, Table 1 gathers the RF chain number, SE, TX/RX complexity (computational complexity for encoding and decoding at the transmit/receive side), diversity order, CSI requirement, and key advantages of the schemes. One can conclude from the table that IM improves the system EE in terms of the hardware cost, computational complexity, and error performance and increases the SE by exploiting the potential of indexing in all available domains. Please refer to the article written by Wen et al. (2017) for details about IM.

Key Applications

Index modulation is accessed in applications related to several facets, such as massive MIMO, full duplex communication, cooperative communication, high-mobility scenarios, wireless-powered communications, visible light communication, and the Internet of Things.

Cross-References

► Spatial Modulation

References

- Abualhiga R, Haas H (2009) Subcarrier-index modulation OFDM. In: Proceedings of IEEE 20th international symposium on personal indoor mobile radio communication (PIMRC), Tokyo, pp 177–181
- Altalib SA, Ali BM, Siddiq AI (2017) BER performance improvement of V2I communication by using OFDM IM exploiting all subcarrier activation patternser. In: Proceedings of international conference on communication, control, computers electron engineering (ICC-CCEE), Khartoum, pp 1–4
- Basar E, Aygolu U, Panayirci E, Poor HV (2011) Space-time block coded spatial modulation. *IEEE Trans Commun* 59(3):823–832
- Basar E, Aygolu U, Panayirci E, Poor HV (2013) Orthogonal frequency division multiplexing with index modulation. *IEEE Trans Signal Process* 61(22):5536–5549
- Basar E (2016) On multiple-input multiple-output OFDM with index modulation for next generation wireless networks. *IEEE Trans Signal Process* 64(15):3868–3878
- Bian Y, Cheng X, Wen M, Yang L, Poor HV, Jiao B (2015) Differential spatial modulation. *IEEE Trans Veh Technol* 64(7):3262–3268
- Cheng X, Wen M, Yang L, Li Y (2014) Index modulated OFDM with interleaved grouping for V2X communications. In: Proceedings of IEEE conference on intelligent transportation systems (ITSC), Qingdao, pp 1097–1104
- Cheng X, Zhang M, Wen M, Yang L (2018) Index modulation for 5G: striving to do more with less. *IEEE Wirel Commun* 25(2):126–132
- Datta T, Eshwaraiah HS, Chockalingam A (2016) Generalized space-and-frequency index modulation. *IEEE Trans Veh Technol* 65(7):4911–4924
- Di Renzo M, Haas H, Ghayeb A, Sugiura S, Hanzo L (2014) Spatial modulation for generalized MIMO: challenges, opportunities and implementation. *Proc IEEE* 102(1):56–103
- Fan R, Yu Y, Guan Y (2015) Generalization of orthogonal frequency division multiplexing with index modulation. *IEEE Trans Wirel Commun* 14(10):5350–5359
- Frenger PK, Svensson NAB (1999) Parallel combinatory OFDM signaling. *IEEE Trans Commun* 47(4):558–567
- Mesleh RY, Haas H, Sinanovic S, Ahn CW, Yun S (2008) Spatial modulation. *IEEE Trans Veh Technol* 57(4):2228–2241
- Mesleh R, Ikki SS, Aggoune HM (2015) Quadrature spatial modulation. *IEEE Trans Veh Tech* 64(6):2738–2742
- Stavridis A, Renzo MD, Haas H (2016) Performance analysis of multistream receive spatial modulation in the MIMO broadcast channel. *IEEE Trans Wirel Commun* 15(3):1808–1820
- Wang J, Jia S, Song J (2012) Generalised spatial modulation system with multiple active transmit antennas and low complexity detection scheme. *IEEE Trans Wirel Commun* 11(4):1605–1615
- Wen M, Li Y, Cheng X, Yang L (2014) Index modulated OFDM with ICI self-cancellation in underwater acoustic communications. In: Proceeding of IEEE asilomar conference on signals, system, computers, Pacific Grove, pp 338–342
- Wen M, Cheng X, Yang L, Li Y, Cheng X, Ji F (2016) Index modulated OFDM for underwater acoustic communications. *IEEE Commun Mag* 54(5):132–137
- Wen M, Cheng X, Yang L (2017) Index modulation for 5G wireless communications, Springer, Berlin
- Younis A, Serafimovski N, Mesleh R, Haas H (2010) Generalised spatial modulation. In: Proceedings conference

record of the 44th Asilomar conference signals, system Computers, Pacific Grove, pp 1498–1502

Zhang M, Cheng X (2017) Spatial-modulation based wireless-powered communication for achievable rate enhancement. *IEEE Commun Lett* 21(6):1365–1368

Index Modulation for OFDM

Jinho Choi

School of Information Technology, Deakin University, Burwood, Australia

Synonyms

Message-driven frequency hopping; Multicarrier index keying with orthogonal frequency division multiplexing; Subcarrier-index modulation aided orthogonal frequency division multiplexing

Definitions

A modulation technique that can convey multiple information bits through the index set of active subcarriers in addition to conventional signal constellations of active subcarriers in an orthogonal frequency division multiplexing (OFDM) system.

Historical Background

Index modulation (IM) can be seen as a generalization of spatial modulation (SM) (Mesleh et al. 2008) that is proposed for multiple input-multiple output (MIMO) systems. In SM, multiple bits can be transmitted by activating a subset of transmit antennas, where the indices of active antennas represent multiple information bits. If there are n_T transmit antennas and K antennas (among n_T) can be activated at a time, there are $\binom{n_T}{K}$ possible combinations, and the number of bits that can be transmitted is $\lfloor \log_2 \binom{n_T}{K} \rfloor$, where $\lfloor x \rfloor$ represents

the largest integer that is smaller than or equal to x . Furthermore, a conventional modulation can also be employed for each active antenna to transmit more bits. In Abu-alhiga and Haas (2009) and Basar et al. (2013), the concept of SM is applied to OFDM, which is called OFDM with IM or OFDM-IM. In OFDM-IM, a subset of subcarriers are active, and their indices are used to convey information. The main advantage of OFDM-IM over conventional OFDM is the energy efficiency and robustness against inter-carrier interference (ICI) as a fraction of subcarriers are active (Xiao et al. 2014). Extensions of OFDM-IM are discussed in Fan et al. (2015) where the number of active subcarriers for each sub-block is not fixed, but variable to increase the number of bits to be transmitted by a sub-block. In addition, another scheme that independently applies IM to in-phase and quadrature phase components of complex symbols is considered to increase the number of bits per sub-block.

A similar idea to OFDM-IM is message-driven frequency hopping (MDFH) (Ling and Li 2009), which might be seen as a generalization of frequency shift keying (FSK). From this point of view, the root of OFDM-IM can also be found in FSK.

Foundations

OFDM-IM is a modulation technique that conveys information by the index set of active subcarriers in addition to conventional modulation for active subcarriers. If the number of subcarriers is N and K of them are active at a time, the number of bits that can be represented by the set of active subcarriers is $\lfloor \log_2 \binom{N}{K} \rfloor + K \log_2 M$, where M is the size of constellation of conventional modulation for active subcarriers. For example, if 8-phase shift keying (PSK) is used as conventional modulation for active subcarriers, and $(N, K) = (8, 2)$, the number of bits transmitted by OFDM-IM becomes $\lfloor \log_2 \binom{8}{2} \rfloor + 2 \log_2 8 = 10$ bits. For comparison, if we consider conventional OFDM where all subcarriers are active with 8-PSK modulation, the

number of bits is $8 \log_2 8 = 24$ bits. Although conventional OFDM can transmit more bits than OFDM-IM (by a factor of $24/10 = 2.4$), the symbol energy of OFDM is higher than that of OFDM-IM by a factor of $8/2 = 4$. Thus, OFDM-IM could be more energy efficient than OFDM.

Compared to SM, OFDM-IM can be more flexible as the number of subcarriers, denoted by N_F , can be large (e.g., $N_F = 512, 1024$, or 2048). However, if N_F is large, there might be too many combinations for the set of active subcarriers, which makes the implementation of OFDM-IM difficult in terms of detection. Thus, subcarriers are divided into multiple sub-blocks, where the size of sub-block is usually small (e.g., 4, 8, or 16) and OFDM-IM is independently performed for each sub-block (Basar et al. 2013).

The diversity order of OFDM-IM under Rayleigh fading is limited (Basar et al. 2013; Ko 2014), which is inherited by OFDM. In OFDM-IM, a data symbol through an active subcarrier experiences fading, and the detection performance of the data symbol is limited by deep fading. In order to increase the diversity gain, the notion of space-time coding can be employed as in Basar (2015b), where in-phase and quadrature phase components of an active subcarrier can come from different symbols with constellation rotation (Khan and Rajan 2006). Channel coding helps improve the diversity gain with a transmit diversity technique (Choi 2017). In general, there are multiple sub-blocks, and each sub-block can be seen as a modulated symbol (by IM). Then, the notion of trellis-coded modulation (TCM) (Ungerboeck 1982) can be employed to increase the signal distance over a sequence of IM symbols with a channel encoder. As a result, the performance can be improved by coding gain as well as diversity gain (Choi and Ko 2018).

As conventional OFDM, OFDM-IM can be applied to MIMO systems. For example, each antenna can transmit independently modulated signals as in Basar (2015a). It is also possible to combine OFDM-IM with SM as in Datta et al. (2016), which can be seen as IM in the frequency

and space domains. Therefore, there can be a number of possible ways to implement MIMO-OFDM-IM.

In OFDM, due to Doppler effect and carrier frequency offset (CFO), the orthogonality of subcarriers can be broken, which results in ICI and degrades the performance. While all the subcarriers are active in OFDM, a fraction of the subcarriers are active in OFDM-IM. As a result, the performance degradation due to ICI in OFDM-IM is lower than that in OFDM (Xiao et al. 2014). This feature makes OFDM-IM appealing to wireless communications that suffer from high Doppler effect and CFO.

Key Applications

For underwater acoustic communications (UAC), OFDM-IM is considered in Wen et al. (2016) to mitigate channel dispersion. Compared to OFDM, OFDM-IM has a low ICI which makes it suitable for UAC where the performance can be degraded by ICI due to inevitable Doppler effect. Due to the same reason, OFDM-IM can also be used for vehicle-to-vehicle and vehicle-to-infrastructure (V2X) applications (Cheng et al. 2014) and would be a favorable modulation scheme for unmanned aerial vehicle (UAV) communications.

Future Directions

The principles of IM can also be applied to multiple access for multiuser systems. That is, the index set of active subcarriers can be used to differentiate users in a multiuser system as sparse code multiple access (Nikopour and Baligh 2013) and sparse index multiple access (Choi 2015). Thus, efficient ways to apply OFDM-IM to multiple access might be an important issue to be studied.

Space-frequency coding can be studied in the context of MIMO-OFDM-IM. A code word becomes a set of active indices in the space and frequency domains. Careful designs might be

required for codes of good energy and spectral efficiency.

There are also other directions as follows:

- Optical wireless: Optical OFDM is widely studied, and optical OFDM-IM is also studied (Baar and Panayrc 2015). Optical OFDM-IM can be investigated for Li-Fi that uses light for short-range very high-speed communications.
- Precoding and single-carrier systems: Precoding can improve the diversity gain at the cost of increasing receiver's complexity. Thus, through precoding, there might be a trade-off between performance and complexity, which is to be exploited to design a precoded OFDM-IM system with a given complexity constraint. If a precoding matrix is the Fourier transform matrix, the resulting OFDM-IM becomes single-carrier IM that can convey information by the set of active (i.e., non-zero) data symbols in the time domain. This can also be seen as a generalization of pulse-position modulation (PPM). Therefore, precoding can lead to a different level of generalization of IM in the time and frequency domains.
- Receiver design with imperfect channel state information (CSI): Imperfect CSI degrades the performance of receivers and results in an error floor in terms of bit error rate. In order to lower the error floor, low-complexity but near optimal receivers are required.

Cross-References

- ▶ [Principle of OFDM and Multi-carrier Modulations](#)
- ▶ [Spatial Modulation](#)

References

Abu-alhiga R, Haas H (2009) Subcarrier-index modulation OFDM. In: 2009 IEEE 20th international symposium on personal, indoor and mobile radio communications, Tokyo, pp 177–181

- Basar E (2015a) Multiple-input multiple-output OFDM with index modulation. *IEEE Signal Process Lett* 22(12):2259–2263
- Basar E (2015b) OFDM with index modulation using coordinate interleaving. *IEEE Wirel Commun Lett* 4(4):381–384
- Basar E, Aygolu U, Panayirci E, Poor H (2013) Orthogonal frequency division multiplexing with index modulation. *IEEE Trans Signal Process* 61(22):5536–5549
- Baar E, Panayrc E (2015) Optical ofdm with index modulation for visible light communications. In: 2015 4th international workshop on optical wireless communications (IWOW), Istanbul, pp 11–15
- Cheng X, Wen M, Yang L, Li Y (2014) Index modulated OFDM with interleaved grouping for V2X communications. In: 17th international IEEE conference on intelligent transportation systems (ITSC), Qingdao, pp 1097–1104
- Choi J (2015) Sparse index multiple access. In: 2015 IEEE global conference on signal and information processing (GlobalSIP), Orlando, pp 324–327
- Choi J (2017) Coded OFDM-IM with transmit diversity. *IEEE Trans Commun* 65(7):3164–3171
- Choi J, Ko Y (2018) TCM for OFDM-IM. *IEEE Wirel Commun Lett* 7(1):50–53
- Datta T, Eshwaraiah HS, Chockalingam A (2016) Generalized space-and-frequency index modulation. *IEEE Trans Veh Technol* 65(7):4911–4924
- Fan R, Yu YJ, Guan YL (2015) Generalization of orthogonal frequency division multiplexing with index modulation. *IEEE Trans Wirel Commun* 14(10):5350–5359
- Khan MZA, Rajan BS (2006) Single-symbol maximum likelihood decodable linear STBCs. *IEEE Trans Inf Theory* 52(5):2062–2091
- Ko Y (2014) A tight upper bound on bit error rate of joint OFDM and multi-carrier index keying. *IEEE Commun Lett* 18(10):1763–1766
- Ling Q, Li T (2009) Message-driven frequency hopping: design and analysis. *IEEE Trans Wirel Commun* 8(4):1773–1782
- Mesleh RY, Hass H, Sinanovic S, Ahn CW, Yun S (2008) Spatial modulation. *IEEE Trans Veh Technol* 57(4):2228–2241
- Nikopour H, Baligh H (2013) Sparse code multiple access. In: 2013 IEEE 24th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), London, pp 332–336
- Ungerboeck G (1982) Channel coding with multi-level/phase signals. *IEEE Trans Inf Theory* 28(1):55–67
- Wen M, Cheng X, Yang L, Li Y, Cheng X, Ji F (2016) Index modulated OFDM for underwater acoustic communications. *IEEE Commun Mag* 54(5):132–137
- Xiao L, Xu B, Bai H, Xiao Y, Lei X, Li S (2014) Performance evaluation in PAPR and ICI for ISIM-OFDM systems. In: 2014 international workshop on high mobility wireless communications, Beijing, pp 84–88

Indoor Visible Light Communication Networks for Camera-Based Mobile Sensing

Yanqun Tang¹, Siyu Tao², Wei Li³,
Zhengyu Zhu⁴, and Zhiguo Shi⁵

¹School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China

²National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China

³National University of Defense Technology, Changsha, China

⁴Zhengzhou University, Zhengzhou, China

⁵Zhejiang University, Hangzhou, China

Synonyms

Light-emitting diode; Optical camera communication; Visible light communication

Definitions

The camera-based sensor receivers available on today's mobile devices prompt a new direction of research and applications, where visible light communication (VLC) can be combined with mobile computing to realize novel forms of sensing and communication applications. Optical camera communication (OCC) utilizes light-emitting diode (LED), LED arrays, or liquid-crystal display (LCD) as transmitters and employs camera-based sensors assembled in consumer electronic devices, such as smartphone and iPad as receivers to serve as an alternative to the photodiode (PD)- or avalanche photodiode (APD)-based receivers.

Historical Background

As a new emerging technology, VLC has great advantages over traditional radio frequency (RF)

communication in terms of communication security, environmental protection, and broadband high speed and is therefore increasingly attracting attention. In recent years, with the popularization of smartphones equipped with high-definition cameras and displays, a new technology called OCC has come into being. Utilizing the pervasive camera-based sensor as receivers, OCC provides a convenient communication mode interaction with the environment without additional hardware modifications (Pathak et al. 2015; Saha et al. 2015; Grobe et al. 2013; Wang et al. 2017; Haas et al. 2017).

It is well known that the camera-based sensor has a high resolution and can classify multiple spatially separated light sources. It is thus a natural multielement optical receiver. Meanwhile, it has a multicolor filter layout and easily separates the blended multicolor lights. Therefore, it is also a natural multicolor receiver. Moreover, a front optical lens can distinguish the light sources in the camera-based sensor and reduce the channel correlation. These appealing features make the image-sensor-based receiver suitable for imaging optical multiple-input multiple-output (MIMO) settings and are expected to support a huge number of parallel transmissions (Hranilovic and Kschischang 2006).

With these advantages of OCC, there are various potential applications, such as indoor localization (Liu and Pang 2003), intelligent transportation (Yamazato et al. 2014), screen-camera communication (Du et al. 2016), privacy protection (Hu et al. 2013), and unobtrusive communication (Yuan et al. 2012; Woo et al. 2012). Very high accuracy and ability to leverage the existing lighting infrastructure make OCC a good candidate for future indoor positioning applications. Meanwhile, VLC over a handset screen-camera link has become an attractive short-range wireless communication solution due to the high availability of camera-equipped smartphones over the past years, where information can be encoded as a stream of images and displayed in screens of smartphone, laptop, or advertisement boards. Besides, camera-based OCC approach is very attractive

for intelligent transportation, particularly on congested highways where the traffic density and the resulting contention impose challenges on traditional RF communications. Moreover, OCC supports unobtrusive communication imperceptible to human eyes between displays and cameras.

Despite various appealing features mentioned above, there are several factors that make OCC deployment challenging. The transmission speed of an OCC system is inherently limited by camera's frame rate. For most of commercial camera-based sensors, the frame rate is unstable and limited by the manufacturing process and image post-processing. And also the frame rate of commercial camera-based sensor is diverse. The frame rate mismatches between transmitter and camera receiver and thus results in mixed frames or lost frames. Consequently, the synchronization problem arises for the image-sensor-based receiver. Meanwhile, the shot noise in OCC originally modeled by Poisson statistics is signal-dependent. The signal-dependent noise model fundamentally requires complex modifications to signal processing and estimation/detection techniques to ensure optical data transmission and reception. Moreover, perspective distortions, misalignments, lens blur, and ambient light degrade the received signal quality and hence reduce information rate in OCC channels.

Foundations

Considering interference coordination and image reception as the main challenges in indoor OCC systems, the current network schemes will be first investigated. In what follows, for the point-to-point link-level system, modulation methods and synchronization algorithms will be discussed. Then, for the multiple-access-level system, several key issues including VLC cell layout, multiple-user access, and mobility management will be emphasized. Note that OCC emerges as a new form of VLC, thus within the levels of network and multiple-access in OCC systems, the slight differences between them are not to be described in the sequel.

Network Schemes

Indoor VLC technology possesses a broad spectrum, and light source deployments within a room can be dense and arbitrary. Hence, as a wireless communication network consisting of VLC, light fidelity (LiFi) (Haas et al. 2016) is an important means of complement conventional RF communications for improving area traffic capacity and dealing with high-traffic requirements.

VLC involves worse path loss and higher frequency than mm-wave communications do. The coverage of VLC cells may be reduced, leading to the concept of VLC attocells that are smaller than domestic microcells, femtocells, and wireless fidelity (WiFi). Hence, the network schemes of LiFi involve two types, i.e., pure light network (Sewaiwar et al. 2015b) and hybrid network (Jin et al. 2015). The downlink denotes the communication link from the optical transmitter to receiver, while the uplink denotes the backhaul link from the receiver to transmitter. Due to different backhaul infrastructures for uplinks, pure light network adopts VLC or invisible light such as a video streaming network with both infrared and visible light, whereas hybrid network is a combination of various technologies such as RF and free space optical links. Wang et al. (2013) presented a bi-directional VLC network, while red-green light is downlink and blue light is uplink. Li et al. (2015) discussed the resource allocation schemes based on different cooperative cell formations in a hybrid network.

Point-to-Point Link Level

In OCC system, the signals are modulated in intensity, color, or time-frequency domain for pervasive optical light sources (illumination LED, LED arrays, displays, or LCD). While the camera which consists of an imaging lens, a camera-based sensor, and a readout circuit, is used as receiver to detect the signals beyond traditional imaging. Generally, camera-based sensors can be classified into two categories, namely, charge-coupled device (CCD) camera-based sensor with global shutter and complementary metal oxide semiconductor (CMOS) camera-based sensor with rolling shutter. In a CCD camera-based sensor, the

global shutter exposes all pixels per frame simultaneously, while in a CMOS camera-based sensor, the rolling shutter exposes one row/column of pixels at a time and reads out the pixel voltages row by row or column by column. Due to excellent performance/cost trade-off, most personal mobile devices or high-end professional camcorders use CMOS camera-based sensors. Thus, only the CMOS camera-based sensors are discussed in the sequel.

Modulation Methods

Practically, an OCC system design requires consideration of human perception of light source flicker. Usually, the commercial camera's frame rate is lower than the transmission frequency. Thus, undersampling-based modulation methods are proposed to ensure no perceptual intensity fluctuation. Both undersampling frequency shift on-off keying (UFSOOK) (Roberts 2013) and undersampling phase shift on-off keying (UPSOOK) (Luo et al. 2014) modulation methods employ the undersampling technique to sample high-frequency signals and obtain the transmitter's states for data recovery. The UFSOOK is a form of direct current (DC) balanced differential coding, while the UPSOOK is similar to the phase shift keying (PSK), where the mark and space have the same frequency and amplitude, but the phases of the corresponding carrier signal are opposite. However, the spectral efficiencies are 0.5 bps and 1bps for UFSOOK and UPSOOK, respectively. In order to improve the spectral efficiency without increasing the number of transmitters, undersampling phase shift pulse amplitude modulation (UPSPAM) is proposed for high efficiency and non-flicking OCC (Luo et al. 2015).

As mentioned above, the CMOS camera-based sensor is a natural multicolor receiver enabled by a color filter array (Rajagopal et al. 2014), and thus color shift keying (CSK) modulation can be performed in OCC. With the ability of commodity cameras to detect a wide range of colors, it is possible to use high-order constellations with CSK (Hu et al. 2015).

It is known that a non-imaging optical MIMO system provides little diversity gain due to the

ill-conditioned channel matrix. However, optical lens can distinguish the source light images and reduce the channel correlation, which makes the imaging receiver one of the most efficient ways to apply optical MIMO. Specially, spatial orthogonal frequency division multiplexing (spatial OFDM) was proposed for imaging optical MIMO system, which encodes information in the spatial frequency domain. Due to the nonnegative requirement, the two-dimensional spatial DC-biased optical OFDM (SDCO-OFDM) (Katz and Bar-Ness 2015) and spatial asymmetrically clipped optical OFDM (SACO-OFDM) (Mondal et al. 2010) are proposed in an imaging optical MIMO system. Similar to one-dimensional optical OFDM, SDCO-OFDM uses a DC bias to convert the bipolar spatial OFDM signal into unipolar and thus suffers poor power efficiency. While the SACO-OFDM has been shown to be more efficient in terms of electrical as well as optical power, where only the odd spatial frequencies are used to carry data and the resultant bipolar signals are clipped at zero to generate a unipolar signal.

Note that generating the real-valued and positive signals imposes extra loss of at least 50% reduction in spectral efficiency. Moreover, the cyclic prefix further reduces the transmission spectral efficiency. Besides, the OFDM-based transmission has disadvantages of the high peak-to-average power ratio (PAPR) and high side lobe. Thus, WPDM has been proposed as an alternative for OFDM (Wong et al. 1997; Nikookar 2013). As one type of filter bank multicarrier (FBMC), WPDM employs orthogonal wavelet packet functions for symbol modulation. Similarly to OFDM, it provides orthogonality between subcarriers, while the basis functions are wavelet packet functions with finite length in time domain. And also in spatial WPSM system, the two-dimensional inverse discrete wavelet packet transform (IDWPT) and the two-dimensional DWPT are performed at the transmitter and receiver, respectively. The waveform bases for spatial OFDM methods are defined in the complex domain; however, the spatial WPDM bases can be defined either in the real domain or complex

domain, which is determined by the scaling and dilatation filters. Moreover, WPDM belongs to overlapped transformation, yielding better spectral concentration.

For the unobtrusive screen-camera communication systems, there are two kinds of modulation methods based on the space-time domain (Woo et al. 2012; Kays et al. 2017) and transform domain (Yuan et al. 2012; Kim et al. 2015), respectively. The space-time domain method primarily utilizes the equivalent low-pass characteristic and critical flicker frequency (CFF) of human vision. This method superposes information directly in time domain and leverages diversity and multiplexing in the spatial domain. The most common implementation of this method is complementary frame design. The transform domain method mainly exploits different sensitivity of human eyes to spatial frequency, superimposing information on the high-frequency component of spatial frequency to achieve the implicit effect.

Synchronization Algorithms

Although perfect synchronization is difficult in a practical OCC system, the imperfect frame synchronization issues must be addressed to achieve a high throughput. Except the oversampling approach, there exist some other algorithms to tackle the synchronization issue, such as per-line tracking and inter-frame coding (Hu et al. 2013) and rateless coding (Du et al. 2016). The per-line tracking and inter-frame coding synchronization algorithm adopts intra-frame per-line color tracking to decode the mixed frames. In what follows, it employs inter-frame erasure coding and frame-based tracking to recover lost frames and to prevent incorrect frames. Note that in this algorithm, all captured frames containing useful information can contribute to the overall frame recovery and result in reliable information transmission. While for rateless coding, it fits well with the channels in which interruptions occur frequently since the encoded bits of the same original information are generated continuously till the successful transmission is completed.

Especially for the unobtrusive screen-camera communication systems, the camera-based sensor might lose frames or capture mixed frame because of the temporal behaviors of displays and cameras (Shu and Wu 2016). The frame synchronization issues consisting of frame loss, frame fusion, and noise are taken into account in Li et al. (2017). The frame loss detecting algorithm is proposed based on the statistic characteristics of difference frames. Moreover, for the frame loss and frame fusion issues, a reasonable mathematical model is established, and robust frame synchronization compensation algorithms are proposed in Li et al. (2018b). At first, a frame loss detecting algorithm is proposed by analyzing spatial mean characteristic of difference frame absolute value. And then a detecting algorithm based on frame fusion fuzzy regions is proposed by utilizing row mean characteristic of difference frame absolute value and row variance characteristic of difference frame. To consider a configuration with low requirements for a camera where a display refresh rate is equal to the camera frame rate, a mixing factor representing a blend of adjacent frames is introduced in Li et al. (2018a). And also, based on the characteristics of the display-camera link and the simplified model under different imaging directions, a robust complementary frame design that can adapt to different imaging directions is derived.

Multiple-Access Level

VLC networks consider communication and illumination, leading to a key issue on cell deployments in the multiple-access level of indoor OCC systems. The links have directionality and vulnerability characteristics, owing to the straight propagation and common obstruction of visible light, respectively. Therefore, cell deployments in an OCC networking encounter several challenges related to these characteristics that differ from those of either WiFi or femtocell networks, such as illuminant cell layout design, multiple-user access, and mobility management.

Cell Layout

The VLC channel gain is influenced by the distance and illuminative orientation from

the light source to receiver. Deqiang et al. (2007) and Miya and Kajikawa (2010) studied both illuminance and received signal power distributions and determined an optimal base station layout. Chen et al. (2013) introduced the concept of multiple-point joint transmission (JT) into a VLC cellular network. Concurrent data transmission from multiple cooperating optical access point (AP) improved the cell-edge user signal quality. Due to the straight propagation of visible light, Rahaim and Little (2013) considered that it is flexible to adjust the coverage of VLC cells. Cell zooming method (Rahaim and Little 2013) was adapted for VLC to show coverage variability when transmitters are capable of dynamically changing their signal power, such that the requirements of both illumination and communication at cell edges are met.

Multiple-User Access

In VLC networks, the majority of multiple access schemes include two typical types, i.e., orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA). In OMA, it is typically realized by means of frequency, time, or code division. ACO-OFDM multiple access are usually adopted in VLC systems; however unipolar and real-time-domain OFDM signals are difficult to ensure compatibility with LED components (Bawazir et al. 2018).

Compared with OMA, NOMA can exploit the entire bandwidth for the time-frequency utilization of all degrees of freedom to increase spectral efficiency. The basic principle underlying NOMA is the superposition of signals in the power domain. Hence, more power is allocated to the farthest users in order to compensate for the user with poor-quality communication. Marshoud et al. (2015) developed a framework based on VLC channel gain ratio for a downlink NOMA power allocation scheme in an indoor multiple-LED VLC system. Yin et al. (2016) analyzed the performance of outage and capacity in NOMA-VLC networks and discussed the relationship between illuminative angles and VLC network throughput. To discuss the user data rate of NOMA based on indoor VLC channels, Tao et al. (2018) presented an alternative lower bound

to asymptotically fit the system throughput of NOMA-VLC.

Mobility Management

The light radiation pattern of a visible light source with strong directionality confines the coverage area of VLC AP, leading to the inflexible mobility and handover limitation. Due to these actual use demands, mobility management is commonly studied in hybrid RF and VLC networks. Sewaiwar et al. (2015a) proposed a mobility support scheme based on color clusters and the mobility support is efficiently provided using a color filter array at receivers. Wang et al. (2016) presented a hybrid network handover scheme based on fuzzy logic, considering the quality of service (QoS) requirements and indoor mobility of VLC users.

Key Applications

In 2007, visible light communications consortium (VLCC) of Japan presented the VLC system standards JEITA CP-1221 and JEITA CP-1222 (Ergul et al. 2015), including the specification of subcarrier bandwidths. In 2008, infrared data association (IrDA) and VLCC proposed physical layer technical requirements, combining infrared data with VLC (Ito and Matsumoto 2009). In 2010, the European Union performed a home gigabit access (OMEGA) project, which investigated the use of VLC in wireless communication (Langer et al. 2008). In 2011, the Institute of Electrical and Electronics Engineers (IEEE) Standards working groups set the 802.15.7 specification of physical layer and medium access control (MAC) (Roberts et al. 2011; Bawazir et al. 2018). As a revision of IEEE 802.15.7 VLC standard, IEEE 802.15.7r1 incorporates new physical layers to support OCC functionalities and MAC modifications (IEEE-802.15-WPANTM 2018). In 2016, the Chinese Wireless Personal Area Network (CW PAN) published the white paper of VLC standards (CESI 2016), involving mobility management and interference mitigation.

Cross-References

- [Internet of Things](#)
- [Massive MIMO](#)
- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

References

- Bawazir SS, Sofotasios PC, Muhaidat S, Al-Hammadi Y, Karagiannidis GK (2018) Multiple access for visible light communications: research challenges and future trends. *IEEE Access* 6(99):26167–26174
- CESI (2016) Visible light communications standards white paper. China Wireless Personal Area Network (CWPA) working group, <http://www.cesi.cn/cesi/guanwanglanmu/biaozhunhuayanjiu/2016/0226/12407.html>
- Chen C, Tsonev D, Haas H (2013) Joint transmission in indoor visible light communication downlink cellular networks. In: *Globecom Workshops (GC Wkshps)*, 2013 IEEE. IEEE, pp 1127–1132
- Deqiang D, Xizheng K, Linpeng X (2007) An optimal lights layout scheme for visible-light communication system. In: *Electronic measurement and instruments, 2007. ICEMI'07. 8th international conference on*, IEEE, pp 2–189
- Du W, Liando JC, Li M (2016) Softlight: adaptive visible light communication over screen-camera links. In: *Computer communications, IEEE INFOCOM 2016-the 35th annual IEEE international conference on*, IEEE, pp 1–9
- Ergul O, Dinc E, Akan OB (2015) Communicate to illuminate: state-of-the-art and research challenges for visible light communications. *Phys Commun* 17:72–85
- Grobe L, Paraskevopoulos A, Hilt J, Schulz D, Lassak F, Hartlieb F, Kottke C, Jungnickel V, Langer KD (2013) High-speed visible light communication systems. *IEEE Commun Mag* 51(12):60–66
- Haas H, Yin L, Wang Y, Chen C (2016) What is LiFi? *J Lightwave Technol* 34(6):1533–1544
- Haas H, Chen C, O'Brien D (2017) A guide to wireless networking by light. *Prog Quantum Electron* 55:88–111
- Hranilovic S, Kschischang FR (2006) A pixelated MIMO wireless optical communication system. *IEEE J Sel Top Quantum Electron* 12(4):859–874
- Hu W, Gu H, Pu Q (2013) Lightsync: unsynchronized visual communication over screen-camera links. In: *Proceedings of the 19th annual international conference on mobile computing & networking, ACM*, pp 15–26
- Hu P, Pathak PH, Feng X, Fu H, Mohapatra P (2015) Colorbars: increasing data rate of led-to-camera communication using color shift keying. In: *Proceedings of the 11th ACM conference on emerging networking experiments and technologies, ACM*, pp 1–12
- IEEE-80215-WPANTM (2018) The IEEE p802.15.7m short-range optical wireless communications task group. http://www.ieee802.org/15/pub/IEEE%20802_15%20WPAN%2015_7%20Revision%20Task%20Group.html
- Ito S, Matsumoto Y (2009) Visible light communication system combined with visible light ID and IrDA. ITE Technical Report 33:45–49
- Jin F, Zhang R, Hanzo L (2015) Resource allocation under delay-guarantee constraints for heterogeneous visible-light and RF femtocell. *IEEE Trans Wirel Commun* 14(2):1020–1034
- Katz E, Bar-Ness Y (2015) Two-dimensional (2-D) spatial domain modulation methods for unipolar pixelated optical wireless communication systems. *J Lightwave Technol* 33(20):4233–4239
- Kays R, Brauers C, Klein J (2017) Modulation concepts for high-rate display-camera data transmission. In: *Communications (ICC), 2017 IEEE International Conference on*, IEEE, pp 1–6
- Kim BW, Kim HC, Jung SY (2015) Display field communication: fundamental design and performance analysis. *J Lightwave Technol* 33(24):5269–5277
- Langer KD, Grubor J, Bouchet O, Tabach ME, Walewski JW, Randel S, Franke M, Nerreter S, O'Brien DC, Faulkner GE (2008) Optical wireless communications for broadband access in home area networks. In: *Anniversary international conference on transparent optical networks*, pp 149–154
- Li X, Zhang R, Hanzo L (2015) Cooperative load balancing in hybrid visible light communications and WiFi. *IEEE Trans Commun* 63(4):1319–1329
- Li M, Hu Y, Tang Y, Yao X (2017) Frame loss detecting for unobtrusive display camera visible light communication. In: *Signal processing, communications and computing (ICSPCC), 2017 IEEE international conference on*, IEEE, pp 1–6
- Li M, Hu Y, Tang Y, Yao X, Tao S, Meng Y (2018a) Multi-imaging direction design for unobtrusive display-camera visible light communication. In: *Submitted to international conference on wireless communications and signal processing*
- Li M, Hu Y, Yao X, Tang Y, Shen Z (2018b) Frame synchronization compensation algorithm for visible light implicit imaging communication. *Acta Opt Sin* 38(1):68–76
- Liu HS, Pang G (2003) Positioning beacon system using digital camera and LEDs. *IEEE Trans Veh Technol* 52(2):406–419
- Luo P, Ghassemlooy Z, Le Minh H, Tang X, Tsai HM (2014) Undersampled phase shift on-off keying for camera communication. In: *Wireless communications and signal processing (WCSP), 2014 sixth international conference on*, IEEE, pp 1–6
- Luo P, Ghassemlooy Z, Le Minh H, Tsai HM, Tang X (2015) Undersampled-pam with subcarrier modulation

- for camera communications. In: Proc Opto-Electron Commun Conf (OECC), pp 1–3
- Marshoud H, Kapinas VM, Karagiannidis GK, Muhaidat S (2015) Non-orthogonal multiple access for visible light communications. *IEEE Photon Technol Lett* 28(1):51–54
- Miya I, Kajikawa Y (2010) Base station layout support system for indoor visible light communication. In: Communications and information technologies (ISCIT), 2010 international symposium on, IEEE, pp 661–666
- Mondal MRH, Panta KR, Armstrong J (2010) Performance of two dimensional asymmetrically clipped optical OFDM. In: GLOBECOM workshops (GC Wkshps), 2010 IEEE. IEEE, pp 995–999
- Nikookar H (2013) Wavelet radio: adaptive and reconfigurable wireless systems based on wavelets. Cambridge University Press, Cambridge
- Pathak PH, Feng X, Hu P, Mohapatra P (2015) Visible light communication, networking, and sensing: a survey, potential and challenges. *IEEE Commun Surv Tutor* 17(4):2047–2077
- Rahaim M, Little T (2013) SINR analysis and cell zooming with constant illumination for indoor VLC networks. In: Optical wireless communications (IWOW), 2013 2nd international workshop on, IEEE, pp 20–24
- Rajagopal N, Lazik P, Rowe A (2014) Visual light landmarks for mobile devices. In: Information processing in sensor networks, IPSN-14 proceedings of the 13th international symposium on, IEEE, pp 249–260
- Roberts RD (2013) Undersampled frequency shift on-off keying (UFSOOK) for camera communications (camcom). In: Wireless and optical communication conference (WOCC), 2013 22nd, IEEE, pp 645–648
- Roberts RD, Rajagopal S, Lim SK (2011) IEEE 802.15.7 physical layer summary. In: GLOBECOM workshops, pp 772–776
- Saha N, Iftekhar MS, Le NT, Jang YM (2015) Survey on optical camera communications: challenges and opportunities. *IET Optoelectronics* 9(5):172–183
- Sewaiwar A, Tiwari SV, Chung YH (2015a) Mobility support for full-duplex multiuser bidirectional VLC networks. *IEEE Photon J* 7(6):1–9
- Sewaiwar A, Tiwari SV, Chung YH (2015b) Novel user allocation scheme for full duplex multiuser bidirectional Li-Fi network. *Opt Commun* 339:153–156
- Shu X, Wu X (2016) Frame untangling for unobtrusive display-camera visible light communication. In: Proceedings of the 2016 ACM on multimedia conference. ACM, pp 650–654
- Tao S, Yu H, Li Q, Tang Y (2018) Performance analysis of gain ratio power allocation strategies for non-orthogonal multiple access in indoor visible light communication networks. *EURASIP J Wirel Commun Netw* 2018(1):154
- Wang Y, Shao Y, Shang H, Lu X (2013) 875-Mb/s asynchronous bi-directional 64QAM-OFDM SCM-WDM transmission over RGB-LED-based visible light communication system. In: Optical fiber communication conference and exposition and the national fiber optic engineers conference, pp 1–3
- Wang Y, Wu X, Haas H (2016) Fuzzy logic based dynamic handover scheme for indoor Li-Fi and RF hybrid network. In: Communications (ICC), 2016 IEEE international conference on. IEEE, pp 1–6
- Wang Z, Wang Q, Huang W, Xu Z (2017) Visible light communications modulation and signal processing. Wiley, Inc., Hoboken
- Wong KM, Wu J, Davidson TN, Jin Q (1997) Wavelet packet division multiplexing and wavelet packet design under timing error effects. *IEEE Trans Signal Process* 45(12):2877–2890
- Woo G, Lippman A, Raskar R (2012) VR Codes: unobtrusive and active visual codes for interaction by exploiting rolling shutter. In: Mixed and augmented reality (ISMAR), 2012 IEEE international symposium on. IEEE, pp 59–64
- Yamazato T, Takai I, Okada H, Fujii T, Yendo T, Arai S, Andoh M, Harada T, Yasutomi K, Kagawa K, et al (2014) Image-sensor-based visible light communication for automotive applications. *IEEE Commun Mag* 52(7):88–97
- Yin L, Popoola WO, Wu X, Haas H (2016) Performance evaluation of non-orthogonal multiple access in visible light communication. *IEEE Trans Commun* 64(12):5162–5175
- Yuan W, Gruteser M, Ashok A, Mandayam N, Dana K (2012) Dynamic and invisible messaging for visual MIMO. In: Applications of computer vision, IEEE workshop on (WACV), vol 00, pp 345–352. www.doi.ieee.computersociety.org/10.1109/WACV.2012.6162992

Industrial Big Data

Lin Gu

SCTS, CGCL, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

Definition

Industrial big data refers to the large amount of data generated by various industrial equipments in production and marketing process. The industrial big data is more orderly, structured, timely correlated and therefore more ready for intelligent analytics to support decision making,

reduce maintenance cost and improve customer experience.

Historical Background

The concept of industrial big data is proposed along with the strategies of the “Industry 4.0” of Germany” in 2012. With the development of “Industrial Internet” in US, the “Manufacturing White Book of 2014” in Japan, and the “Made in China 2025” and “Internet plus” in China, big data based intelligent manufacturing has attracted an ever-growing attention of industrial enterprises and industry related communities globally.

Industrial Big Data Life Cycle

A typical industrial big data life cycle usually consists of the following four parts:

- **Data Generation**

Industrial big data is usually generated by diverse sources in manufacturing spaces and production process, mainly including the following: (1) device data, data generated by the physical devices such as actuators, sensors, and cameras; (2) production data, data generated during the production process, including product design and manufacturing, quality testing, and device maintenance; (3) operation data, data related to the enterprise operation, e.g., enterprise organization, business management, marketing, quality control, and e-commerce plan; and (4) manufacturing chain data: data related to customers, suppliers, business partners, and so on

- **Data Acquisition**

With the development of intelligent manufacturing, the industrial data and manufacturing devices are rapidly growing in both quantity and complexity, requiring industrial data acquisition in real time. Two key technologies as data fusion and data integration are usually applied to synthesize industrial data from

different sources to find their interlink and provide users with a unified view, producing more useful and meaningful information for further analytics and queries.

- **Data Storage**

To manage massive heterogeneous industrial data and enable real-time analytics, the data storage system should be designed for high capacity, low latency, rapid access, and fault tolerance. Big data storage supports storage and I/O operation of massive files and provides basic distributed database functionalities. The storage infrastructure is usually made up of direct attached storage and network attached storage, enabling fast access and retrieval of large amount of data. Typical big data storage architecture/infrastructure solutions include Hadoop, Cassandra, NoSQL, and so on.

- **Data Analytics**

Industrial data analytics is considered as a useful tool to understand customer demands, market trends, and device maintenance, helping to make future plans for industrial enterprises. Due to the large quantity and complex structure of big data, industrial data analytics should be able to provide multi-view understanding of massive data from diverse sources and discover valuable information. Some key technologies include association rule learning, classification tree analysis, genetic algorithms, machine learning, regression analysis, and sentiment analysis.

Enabling Technologies

Modern smart industry can be viewed as a manufacturing cyber-physical system (CPS) consists of both physical devices such as machines, sensors, and information systems to collect, process, and store data produced by these devices. Industrial big data leverages various information technologies such as wireless communications, edge computing, and machine learning for supply chain improvement, device health monitoring, and customer relation management toward low-cost, efficient, and sustainable industry. Many

related enabling technologies have been proposed and applied to the four phases of industrial big data life cycle.

- **Wireless Communications**

Industrial big data are usually acquired via wireless communication from both industrial devices (e.g., machine temperature and vibration) and production procedure (e.g., raw material supply and product quality). Previous wireless communication standards such as LTE-A, WiMax, UMTS, and LTE are usually designed to achieve high throughput and high reliability. However, with the ever-growing sensors and devices involved in the smart factory, it is widely agreed that 5G technology (Palattella et al. 2016) which supports vast device connectivity with higher data rate, higher system capacity, and lower energy consumption should be applied toward high efficient data transmission and massive connections.

- **Edge Computing**

Edge computing offers an ideal platform to offload applications, data, and services from central cloud to local information devices such as servers and base stations (Shi et al. 2016). It enables data processing, analytics, and knowledge generation near the data source with lower latency and higher efficiency. Such advantages fit the needs of smart industry that requires fast data processing and prompt response. For example, one of the key issues in smart industry is machine failure prediction. A large volume of sensing data should be processed in a real-time way, and immediate decision or predictions should be made for the production procedure. In this case, cloud computing is no longer desired with the consideration of high communication cost and latency, while processing the data at edge nodes without data transmission becomes a promising solution.

- **Machine Learning**

Machine learning algorithms and models have been developed for several decades to effectively perform specific tasks without using

explicit instructions or fixed rules. Many different machine learning technologies with different capabilities have been proposed, e.g., decision tree, support vector machine, deep neural network, and reinforcement learning, to perform pattern recognition, trend prediction, and dynamic control (Lei et al. 2016). The industrial big data is usually ordered, structured, timely correlated and therefore ready for further analytics to support smart manufacture. By leveraging machine learning algorithms, enterprises can make the messy industrial data into valuable information, build a knowledge base of the production procedure, and make predictions or decisions automatically.

Key Applications

Through intelligent analytics of large amount of industrial data generated from diverse sources, we can significantly improve industry related management, including marketing decisions and future enterprise resource allocation. Some key applications are listed as follows:

- **Supply Chain Improvement**

Industrial data analytics helps us to predict the future of manufacturing. Thanks to new cloud data management tools and increasing supply chain software platforms, real-time supply chain analytics has become a trend. Specially, enterprise managers now can better understand changing marketing trends and identify investigate risks and thus update supply chain strategies and manage supply chain operations (Wang et al. 2016). For example, data analytics results can be applied to supply network and logistics function design, improve supply chain operations, measure supply chain performance, reduce supply variability, and implement better supply chain strategies at the tactical and operational level.

- **Customer Relation Management**

With the growing of global industrial markets, enterprises are keen to integrate customer-

related data into manufacturing processes with the goal of improving customer satisfaction, calculating investment return on initiatives, and predicting customer behavior (Alqahtani and Saba 2013). The product selling records tell if the sales strategies should be adjusted in one particular area, e.g., keeping low prices or changing advertising plans. Keeping track of the consumers' feedback data or product reviews allows enterprises stay in touch with customers even after the deal. Such data can be analyzed and quantified to help the enterprises to know better about their customers and incorporate necessary changes on the products for better customer experience.

- **Machine Health Monitoring**

The machine condition data and related analytics algorithms to calculate machine health and performance behavior can effectively protect hardware assets and reduce manufacturing downtime (Xu et al. 2015). In modern industrial factories, machine health monitoring is considered as an efficient method to reduce downtime and preserve assets. Usually, some fixed sensors are involved in order to collect data of machine working conditions, e.g., temperature, noise, and vibration frequency. The sensing data could be used for real-time analysis on different machine performance evaluation and future maintenance scheduling.

- **Decision-Making Support**

The main objective of manufacturing decision is to achieve effective and efficient production. The rich production data can be used to assist the decision-making in production control to achieve high reliability and availability (Lee et al. 2014). For example, by analyzing the production data, useful knowledge can be derived to reduce production uncertainties and utilize finite resources on bottleneck sections of manufacturing system, aiming to efficiently reduce the unexpected downtime and machine failures.

With a real-time data management platform, the data from different industrial sources can be

used to provide big data predictive analytics, improve the flexibility of manufacturing pipelines, and maximize the utilization of enterprise resources.

References

- Alqahtani FA, Saba T (2013) Impact of social networks on customer relation management (CRM) in prospectus of business environment. *J Am Sci* 9(7): 480–486
- Lee J, Kao HA, Yang S (2014) Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia CIRP* 16:3–8. Product services systems and value creation. Proceedings of the 6th CIRP conference on industrial product-service systems
- Lei Y, Jia F, Lin J, Xing S, Ding SX (2016) An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Trans Ind Electron* 63(5):3137–3147
- Palattella MR, Dohler M, Grieco A, Rizzo G, Torsner J, Engel T, Ladid L (2016) Internet of things in the 5g era: enablers, architecture, and business models. *IEEE J Sel Areas Commun* 34(3):510–527
- Shi W, Cao J, Zhang Q, Li Y, Xu L (2016) Edge computing: vision and challenges. *IEEE Internet Things J* 3(5):637–646
- Wang G, Gunasekaran A, Ngai EW, Papadopoulos T (2016) Big data analytics in logistics and supply chain management: certain investigations for research and applications. *Int J Prod Econ* 176:98–110
- Xu M, Cai H, Liang S (2015) Big data and industrial ecology. *J Ind Ecol* 19(2):205–210

Industrial Cyber-physical Systems

► Resource Allocation in Industrial IoT

Information Platform

► Architecture and Data Management for Smart Community Information Platform

Information-Centric Networking: Basic Principle and Architecture

Mingchuan Zhang, Junlong Zhu, and Qingtao Wu
School of Information Engineering,
Henan University of Science and Technology,
Luoyang, China

Synonyms

Future networking; Named data networking

Definition

Information-centric networking (ICN) is an approach for communication in a network that provides accessing named data objects as a first order service.

Historical Background

According to the Cisco Visual Networking Index (VNI) report (Cisco visual networking index: forecast and methodology: 2015–2020 [2014](#)), global IP traffic will reach 194 exabytes per month or 2.3 zettabytes per year by 2021. The content retrieval applications are the most of the traffic, such as delivering multimedia content (e.g., Netflix) and sharing user generated data (e.g., YouTube) (Borst et al. [2010](#)). The widespread Internet multimedia traffic will be 82% of global consumer Internet traffic in 2021. The various of video (e.g., TV, VoD, Internet, and P2P) will be roughly 86% of all global IP traffic (Intel [2016](#)). In the current Internet for applications, some distribution services and technologies are used to address the issue of increasing traffic volume by employing content distribution, replication, and caching in different ways. Hence, P2P networking and CDNs are employed, which are implemented typically as an overlay. Moreover,

the distribution approaches access by name, regardless of origin server location. However, scalable content distribution and information sharing cannot be supported adequately via the current Internet, which is centered on the host-to-host communication model. For this reason, the networking community is proposing novel Internet architectures and paradigms to rectify the drawbacks of the current Internet. Information-centric networking (ICN) is a promising Internet architecture in satisfying the requirements of the future networks. The history of ICN can be dated back to the work of Cheriton and Gritter, who introduce a name-based routing in TRAIID (Cheriton and Gritter [2000a, b](#)). Then, Data-Oriented Network Architecture (DONA) was proposed in 2007 (Koponen et al. [2007](#)), which is the first clean-slate ICN architecture. Subsequently, ICN has attracted researcher interests of many networking communities. In order to investigate ICN, many projects have been funded, including Content Centric Networking (CCN) project, Named Data Networking (NDN) project, [MobilityFirst project](#), [ICE-AR project](#), Publish-Subscribe Internet Technology ([PURSUIT](#)) project, Scalable & Adaptive Internet soLutions ([SAIL](#)), COntent Mediator architecture for content-aware nETworks ([COMET](#)), [CONVERGENCE](#), [ANR Connect project](#), [GreenICN project](#), and [ICN-2020 project](#).

Foundations

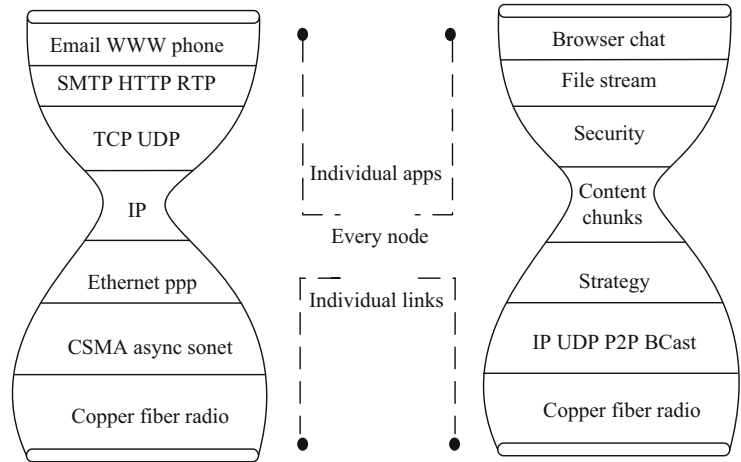
ICN is different from the current Internet, which uses IP addresses to transfer data packets. Moreover, ICN does not care the locations of the contents by assigning contents names. In ICN, each data is assigned a name. In addition, users consume contents by sending requests including content names to the network, which then routes the requests to nearby copies of the desired contents by using the content names. Therefore, names play a key role in forwarding decisions and are used to match requests. Moreover, each ICN node is enabled to cache contents and perform forwarding decisions on a path. Ubiquitous

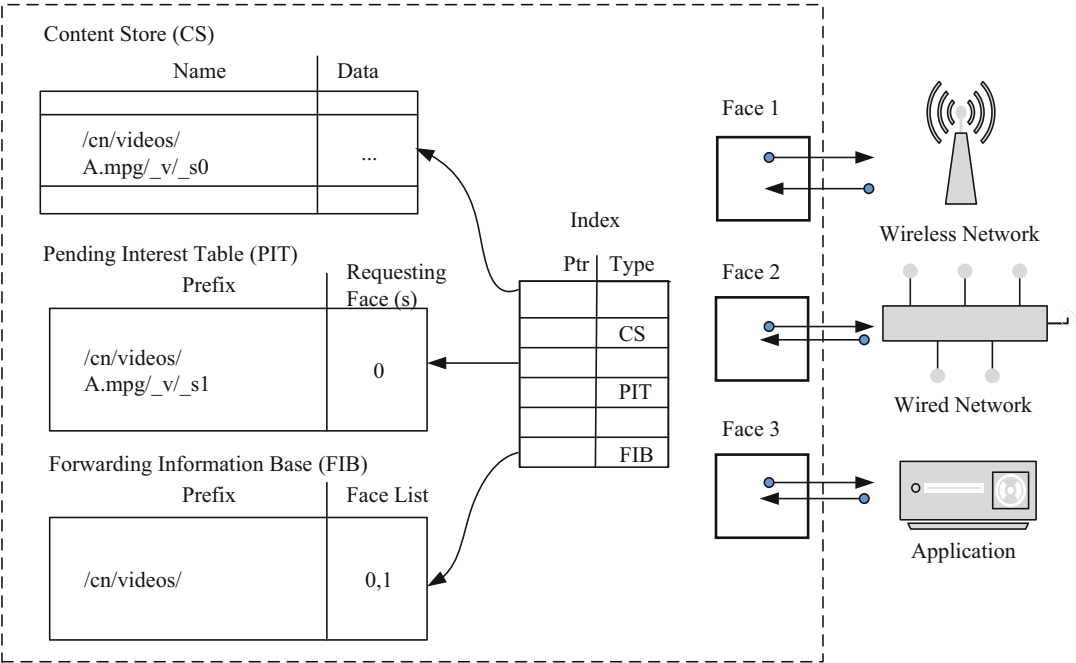
in-network caching can enable sharing and be used to make communication more robust. In addition, transport protocols of ICN are based on a receiver-driver model, i.e., by regulating the request sending rates, message sending rates are controlled by requestors. ICN can achieve retransmission by resending requests. Hence, ICN is a receiver-driven communication model. Compared with the current Internet, the content is decoupled from its location in ICN. Thus, the IP address be replaced by the content name. For example, CCN uses named content chunks to replace the current narrow waist of IP; see Fig. 1. Moreover, ICN users are only interested in what the content is, but do not care where the content comes from.

Next, we describe the ICN forwarding engine, which is shown in Fig. 2. CS is used to cache data, the return path state is stored in the PIT, and probable source(s) for data can be stored in FIB. The routing nodes in the network receive the interest packet when users send a request to the network. Moreover, the routing node looks up its own CS. If CS contains the object, then the content object is sent to the requestor. If CS does not find the object, then the node looks up its own PIT. If no object found, then forward the request toward the data source according to the routing mechanism. In addition, some key functionalities are presented as follows.

- **Naming:** The naming of content is one of the main characteristics of each ICN architectural proposal. For individual content, unique name is required. In all ICN architectures, content names are used to identify contents independently of their locations. In addition, naming schemes are divided into two types: flat and hierarchical namespaces. For example, the hierarchical naming scheme is used to improve routing scalability in Content Centric Networking (CCN). While NetInf uses the flat naming scheme.
- **Security:** The security of ICN is closely related to the naming in several ICN architectures. The current Internet uses encryption and authentication to ensure security. Different from the current Internet, security cannot be bound to the storage location since ICN users can obtain the requested content from multiple sources. In order to verify the integrity of the named content, a security mechanism at receivers is provided by ICN. Moreover, content authentication, access control and authorization, encryption, traffic aggregation and filtering, state overloading, delivering content from replicas, cryptographic robustness, and routing and forwarding information bases are important security features in ICN.
- **Routing and name resolution:** ICN routing constructs a path for finding a named content based on its name, which is provided initially

**Information-Centric
Networking: Basic
Principle
and Architecture, Fig. 1**
The CCN architecture





Information-Centric Networking: Basic Principle and Architecture, Fig. 2 The ICN forwarding engine

- by a requestor. ICN routing is divided into three steps. Firstly, the name of the requested named content is translated to its location by the name resolution step. Secondly, based on its name or location, the request is routed to the content by the discovery step. Finally, the content object is routed to the requestor by the delivery step.
- **Mobility:** ICN natively supports the mobility of the subscriber, since mobile subscribers use content name to request content and the content is transported from publishers to subscribers by a connectionless manner in ICN. Unlike subscriber mobility, publisher mobility support is more difficult, since content identifier and routing locator cannot be separated.
 - **Caching:** Any ICN element is enabled to cache named content, such as end-user devices, proxy caches, and routers. The network utilization and data availability can be improved by content caching, since the content is fetched from nodes that are closer to the requestor in geographical location. Caching approaches can be divided into in-network caching and caching at the network

edge. For in-network caching, the content is cached in the transport network. In general, in-network caching can be divided into off-path and on-path caching. Caching at the network edge means that the content is cached at the edge nodes such as replicated servers.

- **Rate and Congest Control:** Since ICN is a receiver-driver communication model, a requested content can be obtained from multiple sources. Thus, the requestor can control the data rate by regulating the sending rate of its request.

The functionalities above need be considered in all the ICN architectures but rather in the specific ICN architectures.

Key Applications

ICN can be applied to many areas, such as web applications, video streaming and download, and Internet of Things (IoT). The aim is to provide benefits for application developers by employing the advantages of ICN.

Future Directions

From the evolution of ICN, there are some research challenges, including naming schemes for ICN, scalable routing schemes, congestion control, QoS approaches, caching strategies, Metrics for evaluating implementations of ICN, security and privacy, application-protocol design, business, legal and regulatory frameworks, etc.

Cross-References

- [Big Data Networks](#)
- [Software-Defined Wireless Networking](#)
- [Smart Identifier Network](#)

References

- ANR Connect Project. [Online]. <http://anr-connect.org/>
- Borst S, Gupta V, Walid A (2010) Distributed caching algorithms for content distribution networks. In: Proceedings of IEEE INFOCOM'10, Mar 2010, San Diego, pp 1–9
- Cheriton D, Gritter M (2000a) Triad: a scalable deployable NAT-based Internet architecture. Technical report, Jan 2000. [Online]. <http://www-dsg.stanford.edu/triad/>
- Cheriton D, Gritter M (2000b) Triad: a new next-generation Internet architecture, Carnegie Mellon University (CMU), Technical report CMU-CS-11-100, Jan 2000
- Cisco visual networking index: forecast and methodology: 2015–2020 (2014) Technical report, June 2014. [Online]. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html
- Content Centric Networking Project. [Online]. <http://www.ccnx.org/>
- FP7 COMET Project. [Online]. <http://www.comet-project.org/>
- FP7 CONVERGENCE Project. [Online]. <http://www.ictconvergence.eu/>
- FP7 PURSUIT Project. [Online]. <http://www.fp7-pursuit.eu/PursuitWeb/>
- FP7 SAIL Project. [Online]. <http://www.sail-project.eu/>
- GreenICN Project. [Online]. <http://www.greenicn.org/>
- ICE-AR Project. [Online]. <http://ice-ar.named-data.net>
- ICN-2020 Project. [Online]. <http://www.icn2020.org>
- Intel (2016) What happens in an Internet minute in 2016. [Online]. <http://www.visualcapitalist.com/what-happens-internet-minute-2016/>. Accessed 1 Feb 2017
- Koponen T et al (2007) A data-oriented (and beyond) network architecture. ACM SIGCOMM Comput Commun Rev 37(4):181–192

- NSF Mobility First Project. [Online]. <http://mobilityfirst.winlab.rutgers.edu/>
- NSF Named Data Networking Project. [Online]. <http://www.named-data.net/>

Information-Centric Wireless Sensor Networks

Yang Zhang

Wuhan University of Technology, Wuhan, China

Synonyms

[Content-centric wireless sensor networks](#);
[Data-centric wireless sensor networks](#);
[Data-oriented wireless sensor networks](#)

Definitions

An information-centric wireless sensor network is an information-centric network (ICN)-enabled wireless sensor network, where network users aim at requesting and delivering information instead of establishing end-to-end or point-to-point sensor connections and data communication.

Historical Background

A wireless sensor network (WSN) (Akyildiz et al. 2002) is a system containing sensors and related network components for external environment monitoring, information collecting, and transmission, e.g., military surveillance, area detection, and healthcare monitoring. Conventionally, multiple dedicated sensors (e.g., air pollution sensors and temperature sensors) are deployed in physical environments by the sensor owners. A centralized gateway device can be set up for gathering collected data for further process by WSN information and service users. Sensed data need to be merged and routed to corresponding

gateway devices by employing various data delivery algorithms (Al-Karaki and Kamal 2004).

As WSN systems usually provide underlying information delivery services for smart and social applications, it is critical to ensure the quality of information flows within the WSN systems. However, comparing with other popular types of wireless networks, such as mobile ad hoc networks and cellular networks, a WSN system usually involves a large amount of wireless sensor nodes (e.g., over 100 nodes in each 25 m² area (Selavo et al. 2007)), which results in high complexity of network topology as well as end-to-end network node connections. Large overhead of communications may occur when traditional heavyweight networking protocols and communication schemes are applied to WSN systems, e.g., TCP/IP. Additionally, several issues related to information in WSNs are also critical to the applications users of WSNs, e.g., information timeliness, information quality, and fault tolerance/correction issues. To this end, WSN systems need to be viewed and analyzed from a complete different network paradigm perspective.

The concept of information-centric network (Ghodsi et al. 2011) can be applied to WSN systems which mainly focus on efficient data delivery and information service quality, namely, information-centric WSN (Xu et al. 2014; Singh and Al-Turjman 2016; Chen et al. 2016; Al-Turjman 2017). Instead of employing network address based on end-to-end communications (e.g., IP address), an information-centric WSN allows network components to request for information based on the “name” of information. That is, a sensor node sends a request for a particular type of information to the network without being aware of the exact location of the information and the routing details. In a similar manner, a network component who has the information requested by other users only responds to the arriving requests of the information, without establishing an end-to-end route to the request source. Maintaining routing paths for information delivery at the information owner and requester sides is not necessarily required. By this means, information

in the WSN systems is decoupled from physical sources and locations. The delivery processes are implemented by any means of the network itself. Information generation and delivery in WSN systems can be more scalable.

At present, WSN systems have been developed and extended to new scenarios and concepts, e.g., Internet of Things (IoT), which involves smarter sensor devices, as well as cognitive and social network users, including human user network nodes. Social and human factors and behaviors involving information flows, as well as information value and network user incentives in such systems, are extensively studied recently (Mainetti et al. 2011; Gubbi et al. 2013). In this case, Information-centric WSN system architecture can provide an intuitive framework for such studies.

Foundations

System Components and Architecture

A conceptual reference architecture for information-centric WSN systems is shown in Fig. 1, which mainly consists of three layers of different functionalities.

- **Layer I, information sensing and generating layer:** WSN nodes, working in this layer sense, collect and generate raw data either from the environment, e.g., temperature data and video files.
- **Layer II, information processing and networking layer:** Raw data from Layer I is processed and physically delivered in this layer and used to provide services. The services can be formed as structured and interpreted information bundles by selectively combining and employing the raw data passed to this layer, e.g., using machine learning algorithms.
- **Layer III, information application layer:** Information services in this layer are applied to serve the needs and demand of WSN users. WSN users may initiate their requests for different types of information or respond the requests for the information owned.

The information-centric architecture for WSNs allows cross-layer implementations. The key idea is to consider information flows rather than device placements as in conventional WSN systems. This information-centric architecture promotes cross-layer designs and implementations. As many sensor devices have more computing capability nowadays, e.g., smartphones, primitive functions can be provided locally by each device (or a group of devices) to meet immediate information requirements of WSN information and service users. For example, noise can be removed from the sensed information by sophisticated algorithms running by the smartphones. In this regard, sensing devices defined as micro providers (Bohli et al. 2009) can operate across Layers I and II of the ICN architecture as in Fig. 1.

Value of Information

As modern WSN systems involve human and social participants, in information-centric WSN systems, information flow can be valued with the *value of information*. Applying information-centric architecture is intuitive in market-oriented modeling and analysis for information trading in WSN systems. With the architecture, WSN system components can concentrate on the content and timeliness while acquiring data from the system, instead of low-level communication issues. In this case, information is treated as commodities, and information transfer is considered to be market transactions. Additionally, sensing behaviors are considered as services provided by wireless sensor nodes. Activities to handle and process the collected information can be seen as value-adding or value generation processes. Consequently, information becomes expensive in value when the scale of data sensing, collection, and processing and the complexity of WSN applications increase. Therefore, transfers of raw data and processed information should involve value interpretation/proposition and reward implication in terms of either monetary or other types of incentive as in market trading.

Value of information needs to be revealed to WSN system components by means of market-oriented approaches and pricing schemes, such as smart data pricing (Sen et al. 2015), where

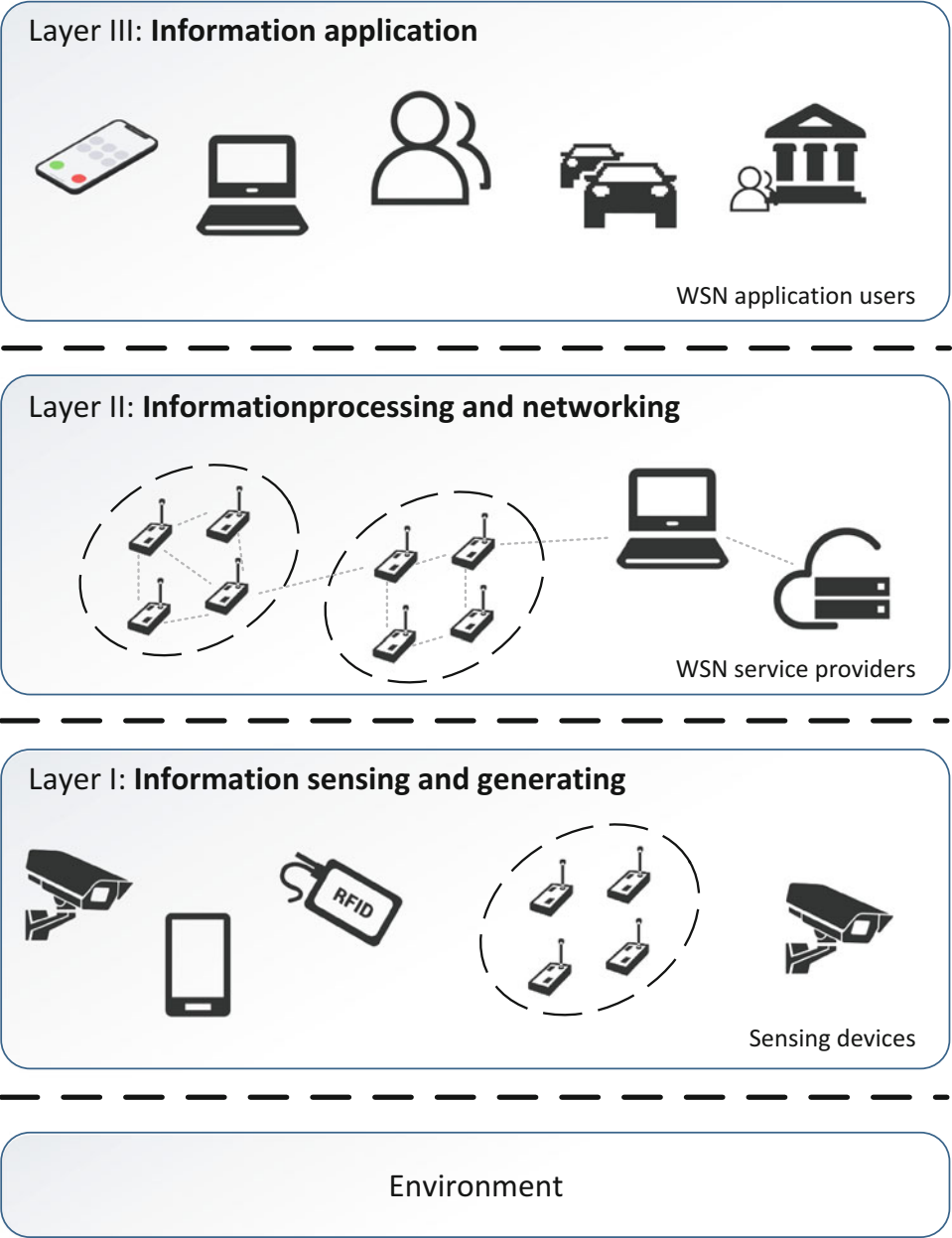
WSN system resource including information and services can be dynamically priced. For example, auction is a classical market-oriented method for valuating and allocating exclusive resources among multiple users. Rational WSN components may negotiate and quote their ask and bid prices for the resources. A practical scenario is a mobile crowdsensing scheme in smart cities (Pouryazdan et al. 2016), where an application task runs a reverse auction to select smartphones (as WSN components) for data sensing and collection. By trading information, market efficiency and social welfare maximization can be achieved.

Sensing as a Service

A set of information trading business frameworks has been categorized into Sensing-as-a-Service (S2aaS) (Charith et al. 2013; Guijarro et al. 2017). WSN nodes perform four types of conceptual tasks in WSN systems (Charith et al. 2013), including data collection, data publishing and transfer, data processing, and data application. A WSN system can be viewed as a WSN market in the S2aaS framework, where information can be circulated as a service-type commodity, and priced based on their market values in the trading processes in terms of their different service qualities.

Age of Information

Age of Information (AoI) (Kosta et al. 2017) can be imported as a concept to measure the quality of information flows in WSN systems in terms of freshness, which is defined as the time interval from the most recent update of the information at the information recipient side. Network users may operate as “free riders” that wait until free or low-priced information is disseminated across the WSN network as time elapses. However, as AoI increases with time, the trading value of information decreases immediately after the information is generated. Discriminatory pricing strategies, i.e., based on the time that the information is made available to a network component, can be adopted. As such, the users should have urgent intent to “buy” information before the information becomes not fresh or even obsolete.



Information-Centric Wireless Sensor Networks, Fig. 1 Conceptual architecture for information-centric wireless sensor networks

Key Applications

Information-centric wireless sensor network paradigms enable the designs and deployments of cognitive and intelligent WSN systems, where

system components mainly value and focus on data requests and data deliveries, instead of end-to-end transmissions. Availability, quality, and valuations of data delivered are important metrics in information-centric wireless sensor networks.

Cross-References

► Data Gathering in Wireless Sensor Networks

References

- Akyildiz I, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. *Comput Netw* 38(4):393–422
- Al-Karaki JN, Kamal AE (2004) Routing techniques in wireless sensor networks: a survey. *IEEE Wirel Commun* 11(6):6–28
- Al-Turjman FM (2017) Information-centric sensor networks for cognitive IoT: an overview. *Ann Telecommun* 72(1):3–18
- Bohli JM, Sorge C, Westhoff D (2009) Initial observations on economics, pricing, and penetration of the internet of things market. *SIGCOMM Comput Commun Rev* 39(2):50–55
- Charith P, Arkady Z, Peter C, Dimitrios G (2013) Sensing as a service model for smart cities supported by internet of things. *Trans Emerg Telecommun Technol* 25(1): 81–93
- Chen B, Liu L, Zhang Z, Yang W, Ma H (2016) BRR-CVR: a collaborative caching strategy for information-centric wireless sensor networks. In: *Proceedings of the 12th international conference on IEEE mobile ad-hoc and sensor networks (MSN)*, pp 31–38
- Ghods A, Shenker S, Koponen T, Singla A, Raghavan B, Wilcox J (2011) Information-centric networking: seeing the forest for the trees. In: *Proceedings of the 10th ACM workshop on hot topics in networks, HotNets-X*. ACM, New York, pp 1:1–1:6
- Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of Things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comput Syst* 29(7):1645–1660
- Guijarro L, Pla V, Vidal JR, Naldi M (2017) Game theoretical analysis of service provision for the Internet of Things based on sensor virtualization. *IEEE J Sel Areas Commun* 35(3):691–706
- Kosta A, Pappas N, Angelakis V (2017) Age of information: a new concept, metric, and tool. *Found Trends Netw* 12(3):162–259
- Mainetti L, Patrono L, Vilei A (2011) Evolution of wireless sensor networks towards the Internet of Things: a survey. In: *Proceedings of the 19th international conference on IEEE software, telecommunications and computer networks (SoftCOM)*, pp 1–6
- Pouryazdan M, Kantarci B, Soyata T, Song H (2016) Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing. *IEEE Access* 4: 529–541
- Selavo L, Wood A, Cao Q, Sookoor T, Liu H, Srinivasan A, Wu Y, Kang W, Stankovic J, Young D, Porter J (2007) Luster: wireless sensor network for environmental research. In: *Proceedings of the 5th international conference on embedded networked sensor systems, SenSys07*. ACM, New York, pp 103–116
- Sen S, Joe-Wong C, Ha S, Chiang M (2015) Smart data pricing: using economics to manage network congestion. *Commun ACM* 58(12):86–93
- Singh GT, Al-Turjman FM (2016) A data delivery framework for cognitive information-centric sensor networks in smart outdoor monitoring. *Comput Commun* 74: 38–51. *Current and Future Architectures, Protocols, and Services for the Internet of Things*
- Xu G, Ngai ECH, Liu J (2014) Information-centric collaborative data collection for mobile devices in wireless sensor networks. In: *Proceedings of the 2014 IEEE international conference on communications (ICC)*, pp 36–41

Infrastructure Design of Data Centers

► Data Center Networking (DCN): Infrastructure and Operation

Infrastructure Sharing, Wi-Fi Sharing

► Sharing Economics

In-Network Caching

► Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks

Innovative Networking Architecture

► Smart Identifier Network

Insider, Internal

► Wireless Sensor Network Security

Insufficient CP-OFDM/OFDMA Systems

- [Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation](#)

Integrated System of Networking, Caching, and Computing

Chenmeng Wang¹ and Richard Fei Yu^{2,3}

¹Chongqing University of Posts and Telecommunications, Chongqing, China

²Carleton University, Ottawa, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

[Caching and computing; Integration of networking](#)

Definition

Integrated systems of networking, caching, and computing refer to integrated frameworks and technologies in wireless or mobile communication systems that take into consideration at least two aspects of networking, caching, and computing, aiming at providing coherent and sustainable services to fulfill various mobile network user demands, such as high-data-rate communication, low-latency information retrieval, and extensive computation resources.

Historical Background

Software-Defined Networking (SDN), Network Functions Virtualization (NFV) (Liang and Yu 2014), and Heterogeneous Network (HetNet) are key supporting technologies from the networking vector of the integrated system. SDN can enable the separation between the data plane and

the control plane, therefore realizing the independence of the data plane capacity from the control plane resource, which means high data rates can be achieved without incurring overhead on the control plane. Meanwhile, the separation of the data and control planes confers high programmability, low-complexity management, convenient configuration, and easy troubleshooting on the network. NFV presents the technology that decouples the services from the network infrastructures that provide them, therefore maximizing the utilization of network infrastructure by offering the possibility that services from different service providers can share the same network infrastructure (Liang and Yu 2015). Furthermore, easy migration of new technology in conjunction with legacy technologies in the same network can be realized by isolating part of the network infrastructure (Liang and Yu 2014). On the other hand, HetNet has the potential to considerably improve network coverage and link capacity (Xie et al. 2012). By employing different sizes of small cells and various radio access technologies in one macro cell, energy efficiency (EE) and spectral efficiency (SE) are significantly enhanced in scenarios such as shopping malls, stadiums, and subway stations (Ma et al. 2004; Yu and Krishnamurthy 2007).

The early stage of caching technology in wireless communication system includes Content Delivery Networks (CDNs) and Peer-to-Peer (P2P) networks, which are proposed as first attempts to confer content distribution and retrieval capabilities on the networks by utilizing the current storage and processing network infrastructure (He et al. 2017b). CDNs employ clusters of servers among the Internet infrastructures and serve the user equipment (UE) with the replications of content that have been cached in those servers. The content requests generated by UE are transmitted through the domain name servers (DNSs) of the CDN to the nearest CDN servers that hold the requested content, in order to minimize latency. P2P networks rely on the storage and forwarding of replications of content by UE. Each piece of UE can act as a caching server in P2P networks. In P2P networks, the sources of content are called

peers. Instead of caching a complete content, each peer may store only a part of the content, which is called a *chunk*. Thus, the content request of a piece of UE is resolved and directed to several peers by a directory server. Each peer will provide a part of the requested content upon receipt of the request. Serving as an alternative to CDN and P2P networking, *Information-Centric Networking (ICN)* emphasizes information dissemination and content retrieval by operating a common protocol in a network layer, which can utilize current storage and processing network infrastructures to cache content.

Powered by network slicing, SDN (He et al. 2017a), and NFV, Cloud Computing (CC) functionalities are incorporated into mobile communication systems, bringing up the concept of *Mobile Cloud Computing (MCC)*, which leverages the rich computation resources in the cloud for user computation task execution, by enabling computation offloading through wireless cellular networks. After computation offloading, the cloud server will create a clone for each piece of UE individually, and then the computation task of the UE will be performed by the clone on behalf of that UE. Nevertheless, due to the fact that the cloud is usually distant from mobile devices, the low-latency requirements of some latency-sensitive (real-time) applications may not be fulfilled by cloud computing (Wang et al. 2018). Moreover, migration of a large amount of computation tasks over a long distance is sometimes infeasible and uneconomical. To tackle this issue, *fog computing* has been proposed to provide UE with proximity to resourceful computation servers. It is a distributed computing paradigm in which network entities with different computation and storage abilities and various hierarchical levels are placed within a short distance from the cellular wireless access network, connecting user devices to the cloud or Internet. A new MCC paradigm similar to fog computing, named *Mobile Edge Computing (MEC)*, is proposed. Being placed at the edge of radio access networks, MEC servers can provide sufficient computation resources in physical proximity to UE, which guarantees the fulfilment of the demand of fast interactive response by low-latency connections (Wang et al. 2017a,b).

Key Points

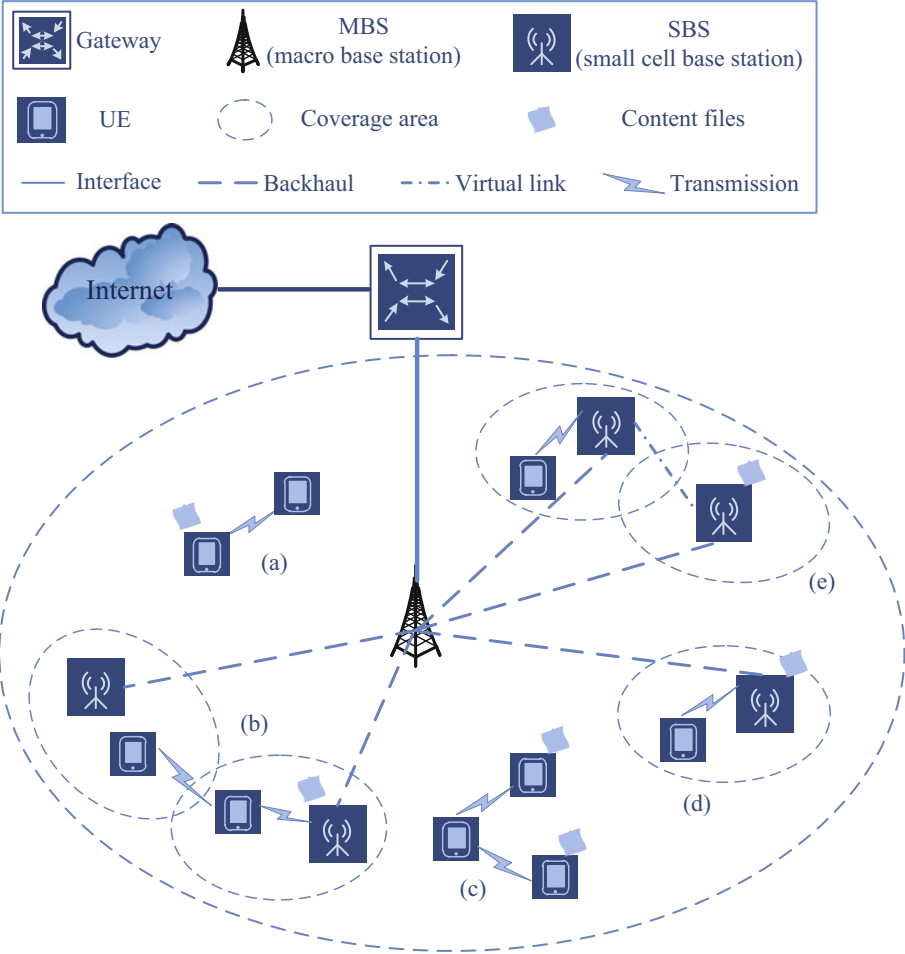
Caching-Networking

The multilayer caching-networking framework is illustrated in Fig. 1; this framework consists of the following five cached content delivery approaches.

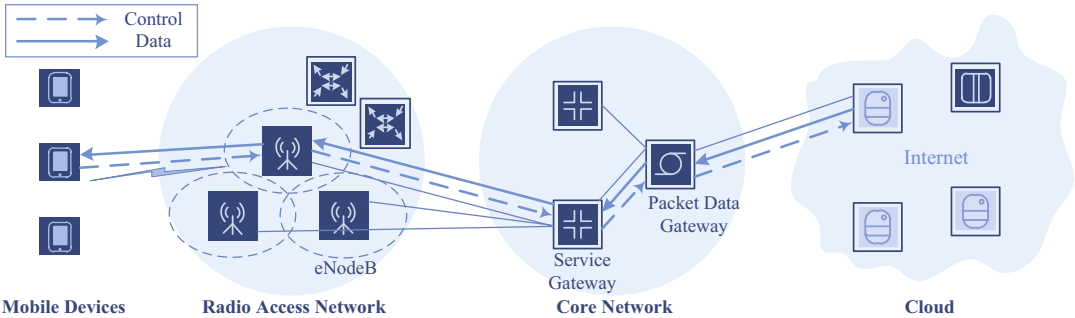
- Device-to-Device (D2D) delivery (Fig. 1a)
Content requested by UE can be retrieved through D2D links from neighboring UE. A high density of UE can contribute to a high probability of successful retrieval of requested content.
- Multihop delivery via D2D relay (Fig. 1b)
D2D communication links can serve as a relay for content delivery from other sources to the destination UE. This approach enables a broader range of caching and delivery.
- Cooperative D2D delivery (Fig. 1c)
If multiple UE hold the content requested by another UE, they can cooperatively transmit the content to the destination UE through a D2D multiple-input multiple-output (MIMO) technology to accelerate the transmission process.
- Direct small cell base station (SBS) delivery (Fig. 1d)
If the requested content was cached in the associated SBS, the SBS can transmit it directly to the destination UE.
- Cooperative SBS delivery (Fig. 1e)
If the content requested by certain UE was stored neither in its nearby UE nor in its associated SBS, it may retrieve and demand the content from an SBS in a neighboring small cell. In this case, the associated SBS will require the demanded content from neighboring SBSs through virtual links, which means these two SBSs are able to communicate with each other via their respective backhauled to the macro cell base station (MBS).

Computing-Networking

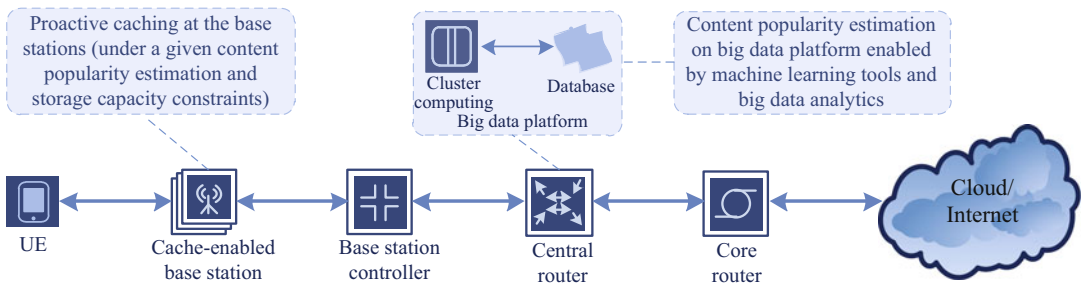
Cloud Mobile Media (CMM) is a typical computing-networking approach and is hence used to illustrate the structure of this type of integrated systems. The overall architecture of CMM is depicted in Fig. 2 (Wang and Dey 2013).



Integrated System of Networking, Caching, and Computing, Fig. 1 Networking-caching framework (Sheng et al. 2016)



Integrated System of Networking, Caching, and Computing, Fig. 2 Cloud mobile media framework, demonstrating data and control flows (Wang and Dey 2013)



Integrated System of Networking, Caching, and Computing, Fig. 3 Caching-computing architecture (Zeydan et al. 2016)

The control commands are transmitted uplink via cellular radio access networks (RAN) to gateways located in core networks (CN) and eventually to the serving clouds. After that, the clouds perform data processing or content retrieval using computing or caching resources in cloud servers, in accordance with user control commands. Then the results are sent back to user devices downlink through CN and RAN.

Caching-Computing

In the context of big data, which has attracted great attention in recent years, data management/processing and data caching have emerged as two key issues in this field. The study of Zeydan et al. (2016) proposes a caching-computing framework in which cached content is moved from the cloud to the RAN edge and a big data platform powered by machine learning and big data analytics is deployed in core site for content selection. As illustrated in Fig. 3, the caching platform at small cell base stations (SBSs) and the computation and execution of the content prediction algorithms at the core site are two key components of the proposed architecture. A big data platform is employed at the core site to perform tracking and prediction of users' demands and behavior for content selection decisions, and SBSs in the RAN are equipped with caching units to store the strategic content selected by the big data platform.

Caching-Computing-Networking

1. Networking-Caching-Computing Convergence

The study in Huo et al. (2016) proposes a framework incorporating networking,

caching, and computing functionalities under the control and management of an SDN control plane. A software-defined approach is employed in the integrated framework, in order to leverage the separation of the control plane and the data plane in SDN, which helps guarantee flexibility of the solution.

2. Networking- and Computing-Assisted Caching

The study in Khan and Ghamri-Doudane (2016) proposes a socially aware vehicular information-centric networking model, in which mobile nodes such as vehicles are equipped with communication, caching, and computing capabilities and they are responsible for computing their eligibility for caching and delivering content to other nodes (i.e., the relative importance of vehicles in the network, with respect to user interests, spatial-temporal availability, and neighborhood connectivity).

Key Applications

Integrated System of Networking, Caching, and Computing is essential in any environment where multiple types of service are requested.

Cross-References

- [Data-Driven Edge Computing](#)
- [Future Wireless Network Architecture and Network Slicing](#)
- [Wireless Edge Caching](#)

References

- He Y, Yu F, Zhao N, Leung VCM, Yin H (2017a) Software-defined networks with mobile edge computing and caching for smart cities: a big data deep reinforcement learning approach. *IEEE Commun Mag* 55(12):31–37
- He Y, Zhang Z, Yu F, Zhao N, Yin H, Leung V, Zhang Y (2017b) Deep reinforcement learning-based optimization for cache-enabled opportunistic interference alignment wireless networks. *IEEE Trans Veh Technol* 66(11):10433–10445
- Huo R, Yu F, Huang T, Xie R, Liu J, Leung V, Liu Y (2016) Software defined networking, caching, and computing for green wireless networks. *IEEE Commun Mag* 54(11):185–193
- Khan J, Ghamri-Doudane Y (2016) Saving: socially aware vehicular information-centric networking. *IEEE Commun Mag* 54:100–107
- Liang C, Yu F (2014) Wireless network virtualization: a survey, some research issues and challenges. *IEEE Commun Surv Tutorials* 17(1):358–380
- Liang C, Yu F (2015) Virtual resource allocation in information-centric wireless virtual networks. In: *Proceedings of IEEE international conference on communications (ICC)*, London, UK, pp 3915–3920
- Ma L, Yu F, Leung VCM, Randhawa T (2004) A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wirel Commun* 11(4):44–51
- Sheng M, Xu C, Liu J, Song J, Ma X, Li J (2016) Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges. *IEEE Commun Mag* 54(8):70–76
- Wang S, Dey S (2013) Adaptive mobile cloud computing to enable rich mobile multimedia applications. *IEEE Trans Multimedia* 15(4):870–883
- Wang C, Liang C, Yu F, Chen Q, Tang L (2017a) Computation offloading and resource allocation in wireless cellular networks with mobile edge computing. *IEEE Trans Wirel Commun* 16(8):4924–4938
- Wang C, Yu F, Liang C, Chen Q, Tang L (2017b) Joint computation offloading and interference management in wireless cellular networks with mobile edge computing. *IEEE Trans Veh Technol* 66(8):7432–7445
- Wang C, He Y, Yu F, Chen Q, Tang L (2018) Integration of networking, caching, and computing in wireless systems: a survey, some research issues, and challenges. *IEEE Commun Surv Tutorials* 20(1):7–38
- Xie R, Yu F, Ji H, Li Y (2012) Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells 11(11):3910–3920
- Yu F, Krishnamurthy V (2007) Optimal joint session admission control in integrated WLAN and CDMA cellular networks with vertical handoff. *IEEE Trans Mobile Comput* 6(1):126–139
- Zeydan E, Bastug E, Bennis M, Kader M, Karatepe I, Er A, Debbah M (2016) Big data caching for networking: moving from cloud to edge. *IEEE Commun Mag* 54(9):36–42

Integration of Networking

- [Integrated System of Networking, Caching, and Computing](#)

Intelligence and Trading with Energy Storage and Renewable Generation of Microgrids

Hongseok Kim

Department of Electronic Engineering, Sogang University, Seoul, South Korea

Synonyms

[Peer-to-peer energy trading of microgrids](#)

Definitions

A microgrid is a power system having at least one distributed energy resource to serve its internal load (Kim 2011). It performs an intentional island operation upon a power disturbance in the electrical distribution network. However, in a normal mode of operation, microgrids are connected to the main grid and can import/export power from/to the power grid or other microgrids.

Historical Background

To set out an agreement for greenhouse gas emission reduction, 195 countries participated in the 2015 United Nations climate change conference (also known as COP 21 or CMP 11) in Paris and established, by consensus, the goal of limiting global warming by below 2 °C. Since the energy sector accounts for roughly two-thirds of greenhouse gas emission, these countries might not effectively combat against climate change without fundamentally shifting

the way of producing energy (e.g., electricity). New York State has recently implemented its Reforming the Energy Vision strategy to create a greener, more resilient, and affordable energy system by utilizing renewable energy sources in community-based microgrids. In the UK, 10 million households are expected to have solar panels on their roofs by 2020 and will be more likely to reformulate electricity market structures as prosumers, that is, both producers and consumers. Although renewable energy can contribute to reducing overall energy costs and carbon emission, its inherent volatility still remains to be resolved as the penetration level of renewable generation continues to increase.

Hence, for reliable operation, microgrids are required to forecast electric loads and renewable generations with reasonable accuracy by utilizing the state-of-the-art forecasting technologies based on machine learning and artificial intelligence. Once determining energy surplus and/or deficit in each hour of operation, microgrids can participate in energy markets, either the existing wholesale electricity market, or the upcoming peer-to-peer (P2P) energy trading market. The focus of this entry is on the emerging P2P energy market.

Electric Load Forecasting

Microgrids first need to forecast their electric loads for reliable operation and energy trading. Thus, electric load forecasting, or simply called load forecasting, is one of the main research areas in the smart grid and can be classified depending on its target forecasting ranges from minutes to years (Ryu et al. 2016). Among these, short-term load forecasting (STLF), which focuses on predicting the load of hours or days ahead, secures attention in the smart grid because of its wide applicability to demand side management, energy storage operation, peak load reduction, etc.

So far, load forecasting is vitally important for the electricity industry in the deregulated electricity market. Furthermore, accurate models for electric load forecasting are also essential for the operation and planning of a utility company.

In addition, demand response is considered as one of the least expensive resources available for operating the system according to the new paradigm under deregulation, and thus, load forecasting is important. Accurate demand forecasting can encourage customers to participate in demand response programs and to receive monetary rewards from the utility company. Meanwhile, in the deregulated retail market, various end-user services, such as customer load management, require load forecasting to attract more customers. These characteristics imply that the role of consumers would be more significant than before, and thus, the scope of forecasting is shifted from the macroscopic view to the microscopic view including the microgrid level.

There are many research activities on forecasting models from classical time series analysis to recent machine learning approaches. Time series analysis, one of the most popular load forecasting methods, includes the autoregressive integrated moving average (ARIMA), exponential smoothing, and the Kalman filter. A double seasonal Holt-Winters model (exponential smoothing) is also popular with better results than ARIMA. Another branch of load forecasting is to use an artificial neural network (ANN) mode. Since ANN can model nonlinearity, it is widely used for various forecasting applications. It is known that network structure plays an important role in ANN because feature information about the model, including the forecasting period or what variables are used, is reflected in the structure of neural networks. The most common type of ANNs is a multilayer perceptron (MLP) that forecasts a load profile using previous load data.

A deep neural network (DNN) is an ANN with more layers than the typical three layers of MLP. The deep structure increases the feature abstraction capability of neural networks. Recent progress of the Internet of Things (IoT) and big data enables the adoption of DNNs in diverse research fields such as computer vision, natural language, signal processing, and deep learning methods are being applied to the domain of load forecasting. Since electricity load is time series data, research studies have been mainly conducted on deep neural network (DNN), recur-

rent neural network (RNN), and long short-term memory (LSTM). DNN is utilized by two different ways (Ryu et al. 2016): one is with pretraining restricted Boltzmann machine and the other is with using rectified linear unit (ReLU) without pretraining. Feed-forward deep neural network and recurrent deep neural network such as LSTM are also possibly combined. Convolutional neural network (CNN) can be also used for load forecasting, and it shows good performance because CNN is good at extracting the features of input data and captures daily, weekly, or seasonal periodicity.

Renewable Energy Forecasting

In addition to load forecasting, renewable energy requires advanced forecasting technologies. Recently, various renewable energy sources such as solar and wind are integrated into the power grid. However, solar power outputs heavily depend on weather conditions such as clouds, temperature, and humidity. This causes substantial uncertainties, which brings adverse effects on economic benefit and the stability of the power grid. Thus, accurate solar forecasting techniques are required to balance the power grid and/or microgrids.

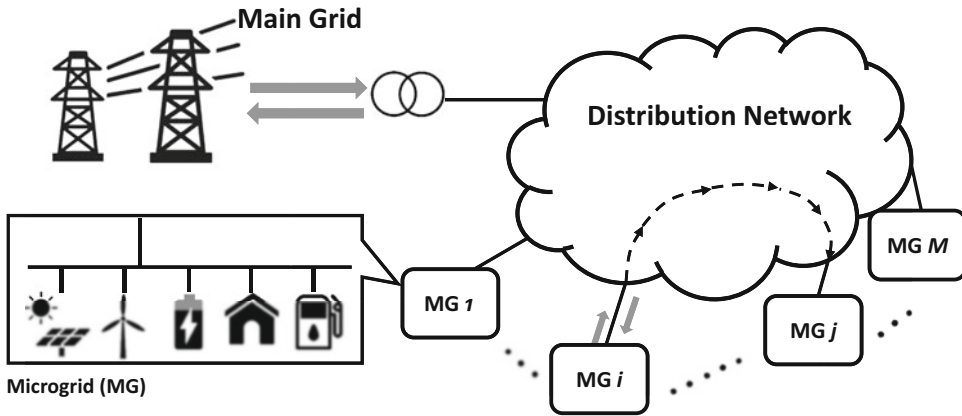
The nonstationary characteristics of solar irradiance mainly come from the presence of clouds because of their stochastic blocking of sun light. This causes the system hard to predict the solar power output. Cloud motion vector can be used for higher accuracy using sky observing cameras, e.g., sky imagers. However, these cameras are usually expensive to be deployed with solar panels, and they are only capable of learning in a very short-term horizon, i.e., less than an hour. Satellite images can be used for generating the solar irradiance forecasting in hourly horizons. However, wide-area satellite images are not suitable for capturing the site-specific information for microgrids. Recent studies have mainly used machine learning approaches to predict multisite solar power output by hourly horizons. There are many time series prediction schemes to extract the relations on historical global (or multisite)

horizontal solar irradiance data and the future solar power output. In general, linear autoregressive prediction models have been used, but nonlinear methods outperform the linear techniques for forecasting horizons more than an hour.

For nonlinear time series prediction, support vector regression (SVR) and ANN are usually used. They are adequate for short-term horizons, i.e., up to 6 h. The numerical weather prediction-based forecasting models are more accurate for long-term horizons more than 6 h. Several studies use forecasted meteorological parameters to improve forecasting accuracy for the short-term forecasting horizons. This meteorological data should be collected by the close weather forecasting station to consider site-specific conditions in a target area. For example, the forecasting accuracy can be improved for 1 h ahead forecasting by using forecasted meteorological data such as clouds, humidity, and temperature. However, this scheme heavily depends on the close weather forecasting station. For forecasting horizons more than an hour, this prediction schemes may be also inadequate because weather forecasting errors are accumulated for longer forecasting horizons. Other studies recently used spatiotemporal historical data for solar prediction to overcome these limitations. Without sky imager or satellite images, spatiotemporal-based prediction scheme traces the indirect showing of cloud image using the spatial information; this means that the spatiotemporal historical data inherently have information about cloud cover and cloud movement.

Energy Storage

No matter how the forecasting of loads and generations is accurate, there are inevitable forecasting errors, which may hinder the reliable operation of power system including microgrids. Thus, it is desirable to have energy storage to buffer the mismatch of prediction and realization (Choi and Kim 2016). The definition of energy storage is stated as “commercially available technology that is capable of absorbing energy, storing it for a period of time, and thereafter dispatching



Intelligence and Trading with Energy Storage and Renewable Generation of Microgrids, Fig. 1 Peer-to-peer energy trading of microgrids

the energy” by the California Renewable Energy Sources Act. Furthermore, having energy storage is good for strategic bidding in energy markets thanks to the flexibility of supplying and absorbing energy. Typically, lithium-ion battery is used for energy storage spanning from tens of kWh for electric vehicles/households to MWh for substation level.

Using energy storage, it is beneficial to charge and discharge energy over appropriate time intervals and encourage microgrids to consume stored energy over peak times, or to shift their usage time for matching energy supply through financial incentive programs, e.g., time-of-use, demand response and peak shift.

In utilizing energy storage, one of the essential factors to consider is battery degradation; due to currently high energy storage prices, it is important to extend battery lifetime as much as possible by properly managing its charging and discharging schedule. If energy storage is designed to schedule without considering battery degradation, battery performance deteriorates rapidly in time, and battery cannot be fully exploited economically. Hence, battery degradation characteristics should be jointly considered in the scheduling process. Note that batteries transform chemical energy to electrical energy (and vice versa) by controlling electron flow between positive and negative electrodes, each of which has different chemical potential. Consequently, battery performance highly depends on chemical

substances in addition to temperature, depth of discharge (DoD), charging and discharging rate, and battery type, etc.

Key Applications: Peer-to-Peer Energy Trading

As discussed so far, intelligent microgrids can predict their electric loads and renewable generations and also know how to leverage energy storage. Then, microgrids can participate in energy markets including P2P energy trading. Specifically, it is desirable to develop a market mechanism for distribution network so that microgrids sell energy to other microgrids in the same distribution network, rather than selling to a utility who suggests a cheaper price. Obviously, P2P energy trading is beneficial to both the sellers and the buyers compared to the trading with the utility because it reduces the redundant intermediate process and the associated profit margin of the operator. However, it is not straightforward to agree on the trading price, which should be indifferent to sellers or buyers. Furthermore, P2P energy trading should be better to achieve social welfare, i.e., minimize the total cost of energy generation usage. If so, regulators have no reason not to support P2P energy trading by providing legitimate support to open a new market (Fig. 1).

Designing a P2P energy trading mechanism considering physical constraints, power flows, and social welfare is still a challenging problem, and the question is *who sells how much energy to whom at what price and why?* Specifically, (1) how to determine a systematic bargaining process in terms of the price, which is not biased toward either sellers or buyers, and the quantity, (2) how to realize a P2P energy trading for networked microgrids even if there is no dedicated distribution line between two microgrids, and (3) how to design a market mechanism which minimizes the total cost considering physical network constraints and power flows in the distribution network.

References

- Choi Y, Kim H (2016) Optimal scheduling of energy storage system for self-sustainable base station operation considering battery wear-out cost. *Energies* 9(6):462
- Kim H (2011) Thottan M. A two-stage market model for microgrid power transactions via aggregators. *Bell Labs Tech J* 16(3):101–107
- Ryu S, Noh J, Kim H (2016) Deep neural network based demand side short term load forecasting. *Energies* 10(1):3

Intelligent Metering System

- [Smart Metering, Specific Challenges, and Solution Approaches](#)

Intelligent Transportation System (ITS)

- [Vehicle Ad-Hoc Networks: Services, Security, and Privacy](#)

Intelligent Transportation Systems

- [Ubiquitous Big Data](#)

Interference Alignment

- [Interference Alignment for Power Line Communications](#)

Interference Alignment for Power Line Communications

Md. Jahidur Rahman¹ and Lutz Lampe²

¹Qualcomm Technologies, Inc., San Diego, CA, USA

²Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada

Synonyms

[Interference alignment](#); [Interference mitigation](#); [Power line communications](#)

Definitions

Power line communication (PLC) refers to a communication method that reuses existing electrical wiring for data transmission between two communication entities. The technique permits the seamless implementation of a communication system without the need for an additional wiring infrastructure. The broadcast nature of data transmission over power lines creates mutual interference between PLC links that can be mitigated through optimized transmission strategies, such as interference alignment (IA). IA allows interfering users to transmit simultaneously by consolidating the space spanned by the interference at receivers within a small number of signaling dimensions while keeping the desired signals separable from interference so that they can be recovered interference-free.

PLC Medium

Data transmission over power lines is an attractive solution for providing communication services, even in hard-to-reach areas through the reuse of existing power grid infrastructures (Lampe et al. 2016). In an in-home electrical network, for example, one can exploit the phase-neutral (P-N) port to enable PLC as a single-input single-output (SISO) system. Besides, in many countries power line networks deploy multiconductor cables, e.g., an additional earth (E) wire, which creates more feeding and receiving possibilities. Therefore, it is also possible to transmit and receive data over multiple conductors and employ spatial processing via well-established multi-input multi-output (MIMO) technology to enable high-data-rate services (see Fig. 1) (Berger et al. 2014).

While at the surface the power line looks to be well-suited for the data transmission, it is not specifically designed for that purpose and is generally considered as a hostile medium for the data communication. With regard to the signal transmission, the PLC channel exhibits high-frequency selectivity due to frequency-dependent power line characteristics and multiple branches in the power line network. Furthermore, owing to changes in network topology or variation of loads with the mains cycle, the PLC channel may also be time varying. This is complemented by a potentially harsh noise environment, where significant noise originates from loads connected to the power line network (Lampe et al. 2016).

Interference in PLC

A closer look at the signal transmission over power lines would reveal that the data communication is essentially an unintended broadcast transmission, since the communication signals can travel through the electricity grid. The broadcast nature of the PLC channel leads to inter-user interference between the feeding and receiving ports that are in close proximity. Therefore, the system performance (e.g., sum-rate) becomes limited by the underlying mutual interference present in the PLC network. As it is illustrated in Fig. 1, simultaneous

pair-wise data communication (i.e., at the same time-frequency resource) between each feeding and receiving port would resemble data transmission and interference scenario in wireless interference networks (Etkin et al. 2008). Another communication protocol is illustrated in Fig. 2, where each feeding port may have signals intended for each of the receiving ports in the network. This is alike the data communication over the wireless X-network (Cadambe and Jafar 2008a).

Noise in PLC

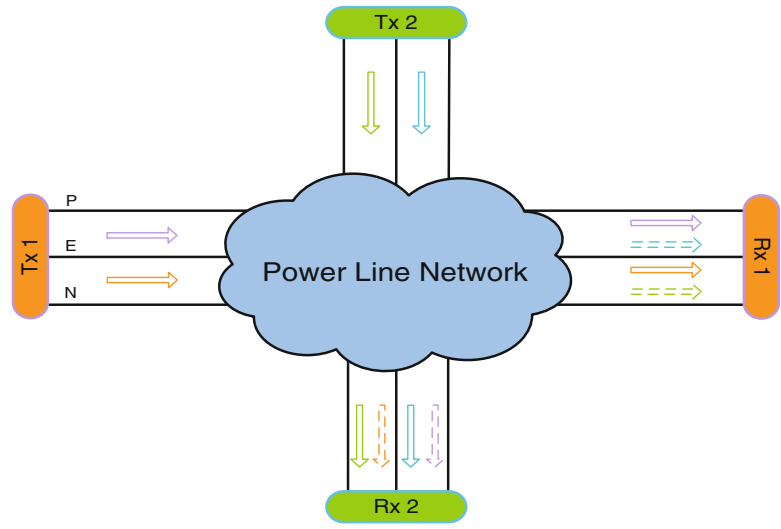
As mentioned above, noise is considered as a major bottleneck for data communication over power lines. It often possesses a richer structure than the additive white Gaussian noise (AWGN) that is generally assumed in many communication system scenarios. PLC noise may have both stationary and nonstationary components (Cano et al. 2016; Cortés et al. 2010). The stationary component, often referred to as the background noise, typically arises from the various components connected to and/or outside of the power line network that can be time varying as well. The nonstationary noise component, which is impulsive in nature, occurs due to power supplies and switching transients. Besides, practical measurements show that the noise across the multiple receiving ports in MIMO PLC may also be spatially colored (Schwager et al. 2012).

Traditional Approaches to Mitigate Interference

Generally, there are three ways to deal with interference in communications networks. First, if the interference is strong, the interfering signal can be decoded, and hence the desired signal can be separated at the receiving port (Sato 1981). Second, when the interference is weak, the interfering signal can be treated as an additional noise (Annapureddy and Veeravalli 2008). On the other hand, when the strength of the interference is comparable to the desired signal, the conventional wisdom is to avoid interference by orthogonalizing the PLC medium either in time

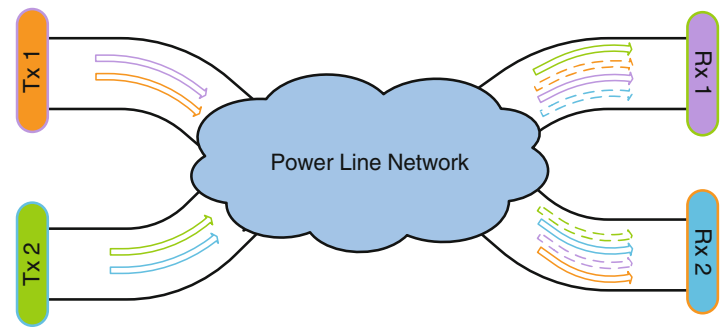
Interference Alignment for Power Line Communications, Fig. 1

An illustration of a MIMO interference network, where each transmitter (Tx 1 and Tx 2) sends two data streams to its paired receiver (Rx 1 and Rx 2, respectively). Solid and dashed arrows denote desired signal and interference, respectively



Interference Alignment for Power Line Communications, Fig. 2

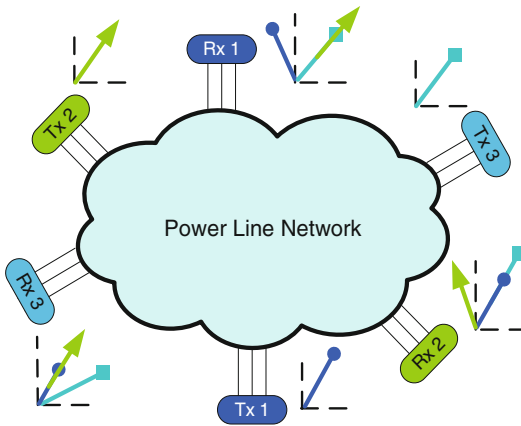
An illustration of a PLC X-network, where each transmitter (Tx 1 and Tx 2) has data to transmit to each of the two receivers (Rx 1 and Rx 2) in the network. Solid and dashed arrows denote desired signal and interference, respectively



or frequency domain. The use of this technique is found in time- or frequency-division multiple-access-based PLC systems (Brown 1998). Therefore, in a K -user PLC network, each user would only get $\frac{1}{K}$ of the total resources per channel use. In such a network where the transmission from any feeding port interferes reception at all other receiving ports, there will be $K - 1$ interfering signals. In general, K signaling dimensions will be needed to recover the desired signal at each receiving port. This poses the fundamental question – is it possible to do any better, i.e., recover the desired signal within a reduced signaling dimension, and if so, how many signaling dimensions are really needed?

Advanced Solution to Mitigate Interference: Interference Alignment

The above question was solved using the notion of interference alignment (IA), which was introduced as an approach to maximize the interference-free signal space for the desired user (Maddah-Ali et al. 2006, 2008; Cadambe and Jafar 2008b). That is to say, if the interference signals could be consolidated into a reduced signaling dimension so that they do not span the entire signal space at the receiver, while at the same time the desired signal could avoid falling into the interference space, it would be possible to recover the desired signal interference-free. This



Interference Alignment for Power Line Communications, Fig. 3 An illustration of the 3-user IA for a 3-conductor cable MIMO PLC interference network achievable via two spatial signaling dimensions. (Illustration adopted from Jafar 2011, Fig. 3.4.)

is done via the technique of IA. The alignment of the interference can be obtained in time domain (e.g., via symbol extension), in frequency domain (e.g., over multiple subchannels), or in spatial domain (e.g., via MIMO transmission). As it is illustrated in Fig. 3 for a MIMO PLC network with three conductors and thus two spatially multiplexed signals, all three users can communicate simultaneously over the spatial domain via two signaling dimensions. At each receiver, two interfering signals are aligned over a signaling dimension, while the desired signal is received interference-free on the second dimension. The idea is considered to be a breakthrough since now interference networks, e.g., the PLC network, can achieve a much higher rate employing this transmission technique (Jafar 2011).

IA for PLC Networks

While IA has been extensively studied in the context of wireless communications, only recently the researchers have shown the benefits of the IA for PLC systems (Rahman and Lampe 2015, 2017, Submitted). In this regard, relevant algorithms of the IA transmission for different PLC

networks and key performance results are presented below.

IA for MIMO PLC Interference Networks

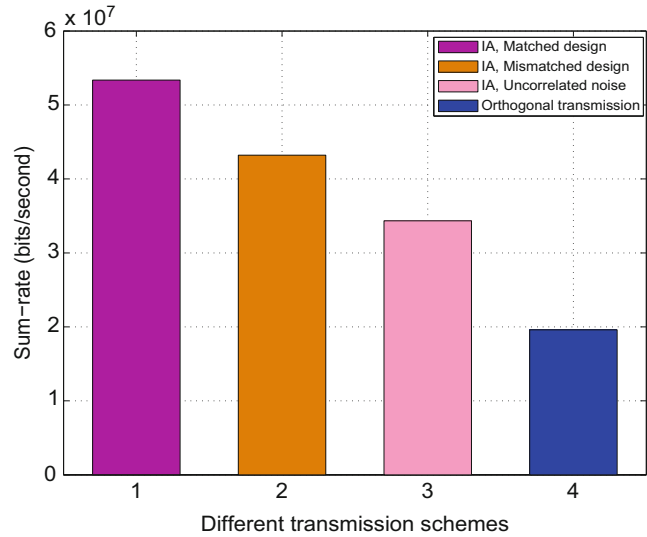
Following the successful application of iterative IA algorithms, such as the minimum interference leakage (Min-IL) and the maximum signal-to-interference-plus-noise ratio (Max-SINR) algorithms in MIMO wireless communications (Gomadam et al. 2011), they were studied in the context of MIMO PLC networks (Rahman and Lampe 2015). The Min-IL algorithm is designed to iteratively minimize the leakage interference power at each receiver in order to find the optimal IA filters. To this end, the quality of IA performance is measured by the power of the leakage interference at each receiver. One giveaway of the Min-IL algorithm is that it only seeks to achieve perfect IA and makes no attempt in maximizing the desired signal power. The Max-SINR algorithm, on the other hand, chooses the IA filters in a way that maximizes the SINR at the receivers, instead of only minimizing the leakage interference (Gomadam et al. 2011).

In Rahman and Lampe (2015), the Min-IL algorithm has been used to confirm the feasibility of the IA for a MIMO PLC interference network as shown in Fig. 3. To that end, the broadband PLC transmission using the frequency band from 2 to 30 MHz and multicarrier modulation using orthogonal frequency-division multiplexing (OFDM) was considered. In the presence of AWGN, the performance results obtained using the Max-SINR algorithm suggest that the sum-rate gain with IA over orthogonal transmission techniques is around 30% for the considered 3-user 2×2 (3-conductor cable) MIMO PLC network. The research findings also indicate some disparity in comparison to the IA performance in wireless communications owing to shared PLC channels.

The study in Rahman and Lampe (2015) has been extended by considering measured MIMO PLC noise and specifically the impact of the spatial noise correlation on the system performance in Rahman and Lampe (2017, Submitted). Two network configurations, i.e., transmission using the P-E/N-E and the P-N/N-E ports, are

Interference Alignment for Power Line Communications, Fig. 4

Comparison of the sum-rate performances among different IA designs and orthogonal transmission across P-E/N-E ports for a noise dataset collected in Xirivella, Spain. The matched IA design outperforms the mismatched IA design assuming uncorrelated noise, as well as the cases of actually uncorrelated noise and orthogonal transmission



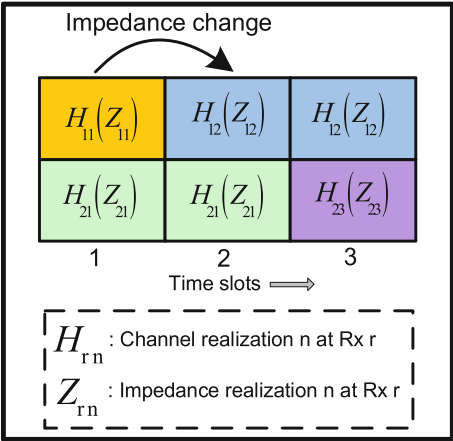
studied. In particular, the study quantified the gains of a matched IA design over that of a mismatched one, where the computation of the IA filters ignores the spatial noise correlation. For a noise dataset from Xirivella, Spain (Schwager et al. 2012), which has an average noise power spectral density of about -74 dBm/Hz, it is found that the average rate gains of the matched over the mismatched IA design are around 30% and 10% across P-E/N-E and P-N/N-E receiving ports with spatial noise correlation coefficients 0.57 and 0.22, respectively (Rahman and Lampe 2017, Submitted). This suggests that the higher the spatial correlation across receiving ports (e.g., across P-E/N-E ports), the larger the rate gains. The study also reveals that the spatial noise correlation actually helps improve the system sum-rate. As shown in Fig. 4, the design considering the spatial noise correlation, i.e., transmission scheme 1, outperforms the mismatched design assuming uncorrelated noise (scheme 2), as well as the cases of actually uncorrelated noise (scheme 3) and orthogonal transmission (scheme 4). The performance benefits hold even when the orthogonal transmission is supplied with 3 times the transmit power, in which case the total transmit power induced into the network is the same as for IA where all three users transmit simultaneously (Rahman and Lampe 2017, Submitted).

IA for SISO PLC X-Networks

Another seemingly feasible idea that was recently attempted to enable IA for the SISO PLC X-network (as shown in Fig. 2) was the technique of blind interference alignment (BIA) (Lampe et al. 2017). The key advantage of a BIA scheme is that it does not require the channel state information be available at the transmitter (Gou et al. 2011; Jafar 2012). However, the BIA scheme requires that the channel seen by each receiving port changes following a specific pattern. The desired patterns of the channel variations over time slots are shown in Fig. 5. It is meant to facilitate the interference cancellation at a given receiving port. One possibility to induce these channel changes is by impedance modulation as suggested in Fig. 5. However, even though the load modulation changes the channel transfer characteristics over time slots, it has been demonstrated based on transmission line modeling of the PLC channel to be ineffective for blind IA (Lampe et al. 2017).

Achievability of the BIA for PLC X-Networks

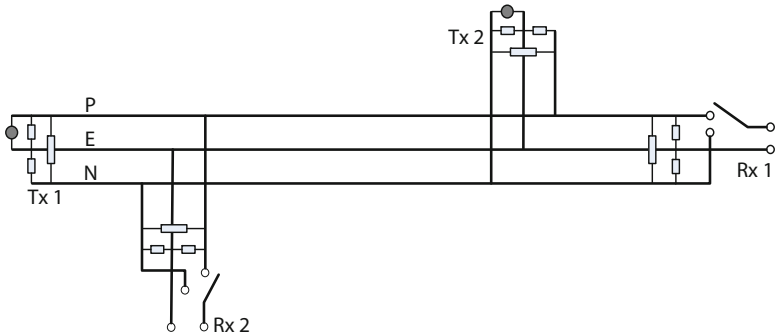
To control the channel response to a desired receiving port while keeping channel responses to other receiver ports unchanged can be achieved via a concept similar to antenna-staggering in the



Interference Alignment for Power Line Communications, Fig. 5 An illustration of the impedance modulation in order to achieve desired channel patterns over three time slots (Illustration adopted from Jafar 2011, Fig. 4.9), which has been proved to be ineffective for the blind interference alignment in PLC networks

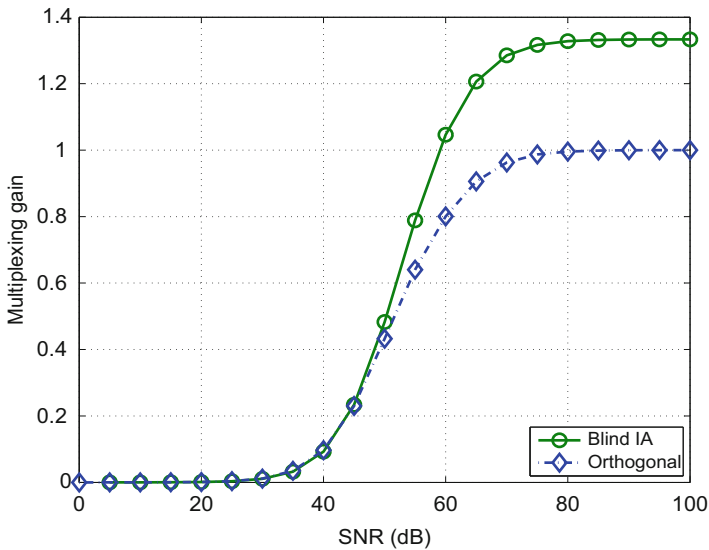
Interference Alignment for Power Line Communications, Fig. 6

An illustration of the transmission scheme for the achievability of blind IA for the PLC X-network using multiple receiving ports



Interference Alignment for Power Line Communications, Fig. 7

A comparison of the multiplexing gains as a function of the transmitter side SNR (on the x-axis) for the proposed BIA and orthogonal transmissions



case of wireless communications. However, this technique requires the use of multiple receiving ports, i.e., a multiconductor structure as in MIMO or SIMO PLC networks. Figure 6 shows such a MIMO PLC X-network with switches at the receiving end to alternate between the receiving ports. The switching between the receiver ports essentially serves the same purpose of antenna switching in wireless communications (Gou et al. 2011). Since switching at a given receiving port has negligible impact on the channel response to the other receiving port due to the attached impedance in parallel as shown in Fig. 6, it can produce the desired channel patterns as illustrated in Fig. 5 for the realization of blind IA. The simulation of the proposed BIA scheme is carried out assuming a distance of 20 m between the nodes. The PLC transmitters and receivers are terminated with load resistances chosen from the

range between 25 and 100 Ω . Each transmitter sends a single data stream intended for each of the receivers in the network. Both the transmitters are provided with the same power and AWGN is assumed. For the evaluation of performance, the multiplexing gain is calculated via computing the slope of total sum-rates as a function of the transmitter side signal-to-noise ratio (SNR). The results in Fig. 7 suggest that a multiplexing gain of 4/3 can be achieved with BIA, which exceeds the multiplexing gain for an orthogonal transmission. The maximum multiplexing gain that is achievable for the considered PLC X-network configuration in Fig. 6 is 4/3 (Gou et al. 2011). This confirms the effectiveness of the BIA with the proposed multiconductor transmission scheme.

Conclusions

While IA has been widely studied in wireless communications, only recently this technique has been investigated in the context of PLC. This chapter presents a synopsis on the IA techniques for PLC networks. To this end, IA techniques are discussed and relevant results are summarized. While IA is shown to improve the sum-rate performance, the seemingly simple realization of blind IA using the impedance modulation is not effective for PLC X-networks. However, a transmission technique similar to antenna-staggering in wireless communications via the use of multiple receiving ports is able to achieve the maximum multiplexing gain for blind IA over this network.

References

- Annapureddy VS, Veeravalli VV (2008) Gaussian interference networks: sum capacity in the low interference regime. In: *Proceeding of the IEEE international symposium on information theory (ISIT)*, pp 255–259
- Berger LT et al (2014) *MIMO power line communications: narrow and broadband standards, EMC, and advanced processing*. CRC Press
- Brown PA (1998) *Powerline communications network employing TDMA, FDMA and/or CDMA*. US Patent 6,144,292, 7 Nov 2000
- Cadambe VR, Jafar SA (2008a) Degrees of freedom of wireless X networks. In: *Proceeding of the IEEE international symposium on information theory (ISIT)*, pp 1268–1272
- Cadambe VR, Jafar SA (2008b) Interference alignment and degrees of freedom of the K -user interference channel. *IEEE Trans Inf Theory* 54(8): 3425–3441
- Cano C et al (2016) State of the art in power line communications: from the applications to the medium. *IEEE J Sel Areas Commun* 34(7):1935–1952
- Cortés JA et al (2010) Analysis of the indoor broadband power-line noise scenario. *IEEE Trans Electromagn Compat* 52(4):849–858
- Etkin RH et al (2008) Gaussian interference channel capacity to within one bit. *IEEE Trans Inf Theory* 54(12):5534–5562
- Gomadam K et al (2011) A distributed numerical approach to interference alignment and application to wireless interference networks. *IEEE Trans Inf Theory* 57(6):3309–3322
- Gou T et al (2011) Aiming perfectly in the dark-blind interference alignment through staggered antenna switching. *IEEE Trans Signal Process* 59(6): 2734–2744
- Jafar SA (2011) *Interference alignment – a new look at signal dimensions in a communication network*. Now Publishers, Delft
- Jafar SA (2012) Blind interference alignment. *IEEE J Sel Topics Signal Process* 6(3):216–227
- Lampe L et al (2016) *Power line communications: principles, standards and applications from multimedia to smart grid*, 2nd edn. Wiley
- Lampe L et al (2017) Characteristics of power line networks: diversity and interference alignment. In: *Proceeding of the IEEE international symposium on power line communications and its applications (ISPLC)*, pp 1–6
- Maddah-Ali MA et al (2006) *Communication over X channel: signaling and multiplexing gain*. University of Waterloo, Technical report, UW-ECE-2006-12
- Maddah-Ali MA et al (2008) Communication over MIMO X channels: interference alignment, decomposition, and performance analysis. *IEEE Trans Inf Theory* 54(8):3457–3470
- Rahman MJ, Lampe L (2015) Interference alignment for MIMO power line communications. In: *Proceeding of the IEEE international symposium on power line communications and its applications (ISPLC)*, pp 71–76
- Rahman MJ, Lampe L (2017, Submitted) Rate improvement in MIMO power line communications through interference alignment
- Sato H (1981) The capacity of the Gaussian channel under strong interference. *IEEE Trans Inf Theory* 27(6): 786–788

Schwager A et al (2012) European MIMO PLT field measurements: overview of the ETSI STF410 campaign and EMI analysis. In: Proceeding of the IEEE international symposium on power line communications and its applications (ISPLC), pp 298–303

Interference Alignment in Wireless Networks

Huacheng Zeng¹, Y. Thomas Hou², and Wenjing Lou²

¹University of Louisville, Louisville, KY, USA

²Virginia Tech, Blacksburg, VA, USA

Synonyms

Cooperative transmitter-side interference management

Definitions

Interference alignment (IA) is a baseband signal processing technique at multiple cooperative transmitters to achieve the directional alignment of interfering signals and preserve the resolvability of desired signals at each receiver.

Historical Background

Interference alignment (IA) is a recent advance in interference management. It offers a new direction to manage the mutual interference in cooperative wireless networks. IA is a transmitter-side signal processing technique. It refers to the construction of baseband signals at multiple cooperative transmitters so that the radio signals are aligned to the same direction at their unintended receivers while remaining resolvable at their intended receivers. Despite similar ideas having been around for many years, the concept of IA was coined by Cadambe and Jafar in a seminar paper

Cadambe and Jafar (2008) in 2008, which showed that by using IA, the K -user interference channel can achieve $K/2$ degrees of freedom (DoFs) regardless of the number of users in the channel. This result is significant as it shows that the DoFs (first-order approximation of capacity) of a wireless interference network can increase linearly with the number of users in the network. This scaling-law result challenged the conventional understanding of “wireless networks are interference limited” and thus attracted a large amount of research attention in wireless communication community. Since then, the concept of IA has been applied to a variety of wireless networks in increasingly sophisticated forms, such as the K -user MIMO interference channel (Bresler et al. 2014), the cellular network (Suh et al. 2011), the MIMO Y channel (Lee et al. 2010), the X network with arbitrary number of users, and the complex interference channel. An information theoretic tutorial of IA in different wireless networks is presented in Jafar (2011). In addition to advances in theory, IA has been studied on wireless testbeds to explore its practicality and performance in real-world wireless systems. This research line has also produced many exciting results on IA (see, e.g., El Ayach et al. 2010 and Gollakota et al. 2009).

Foundations

In wireless networks, based on the domain in which the signals are manipulated at the cooperative transmitters, the results of IA can be divided into different categories: space-domain IA (SD-IA), frequency-domain IA (FD-IA), and time-domain IA (TD-IA).

Space-Domain IA SD-IA refers to a joint construction of baseband signals at the transmitters by precoding their outgoing data streams onto their multiple antennas so that at each receiver, the interfering streams are overlapped in the space domain, while the desired data streams remain resolvable. Since the interfering streams are overlapped at a receiver, the dimension of

the interference subspace is reduced, making it possible to receive more concurrent data streams free of interference at this receiver. Therefore, to achieve SD-IA in a network, each node must have multiple antennas.

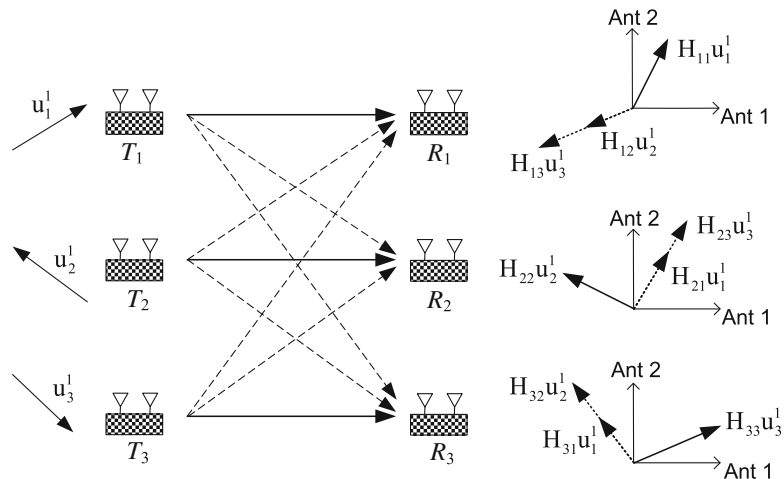
Figure 1 shows an example of SD-IA in a three-link network, where each node has two antennas. Without using SD-IA, this network can transport two data streams from its transmitters to its receivers. In contrast, by using SD-IA, this network can transport three data streams from its transmitters to its receivers. The SD-IA scheme for this network is illustrated in Fig. 1. At transmitter T_i , denote $\mathbf{u}_i^k$ as the precoding vector for the k th stream. Denote $\mathbf{H}_{ji}$ as the channel matrix between receiver R_j and transmitter T_i . Then, SD-IA can be achieved at the receivers through jointly constructing the precoding vectors as follows: $\mathbf{u}_1^1 = \text{eigvec}(\mathbf{H}_{21}^{-1}\mathbf{H}_{23}\mathbf{H}_{13}^{-1}\mathbf{H}_{12}\mathbf{H}_{32}^{-1}\mathbf{H}_{31})$, $\mathbf{u}_2^1 = \mathbf{H}_{32}^{-1}\mathbf{H}_{31}\mathbf{u}_1^1$, $\mathbf{u}_3^1 = \mathbf{H}_{23}^{-1}\mathbf{H}_{21}\mathbf{u}_1^1$, where $\text{eigvec}(\cdot)$ is an eigenvector of a square matrix. It can be verified that by using the above precoding vectors, the interfering streams will be aligned in the same direction at each receiver; and the desired data stream will lie in an independent direction (free of interference) at each receiver, as shown in the figure. Therefore, one data stream can be transported on each link, and a total of three data streams can be transported in this network.

Frequency-Domain IA FD-IA refers to a joint construction of baseband signals at the transmitters by precoding their outgoing data streams onto a set of frequency subcarriers so that at each receiver, its interfering streams are overlapped in the frequency domain, while its desired data streams remain distinguishable. Different from SD-IA, FD-IA does not require the network to have multiple antennas at each node. Instead, it requires the network to have multiple orthogonal frequency channels. As such, FD-IA is particularly suited for OFDM-modulated wireless systems. FD-IA was first studied for cellular networks by Suh and Tse in (2008), in which they proved that a FD-IA scheme could achieve $K/(G\sqrt{K} + 1)^{G-1}$ DoFs for each cell, where G is the number of cells and K is the number of users in a cell. As the number of users in a cell becomes large ($K \rightarrow +\infty$), each cell can achieve one DoF. This means that each cell (base station) can serve its users as if there were no inter-cell interference. In Suh et al. (2011), Suh et al. extended their FD-IA scheme to the downlink of a cellular network and showed that their FD-IA scheme works for the network where feedback information is limited in a cell. In Zeng et al. (2017b), Zeng et al. developed a FD-IA model for a heterogeneous cellular network and proved the feasibility of the FD-IA model.

Figure 2 shows an example of FD-IA in a single-antenna wireless network, which has three orthogonal frequency subcarriers for data trans-

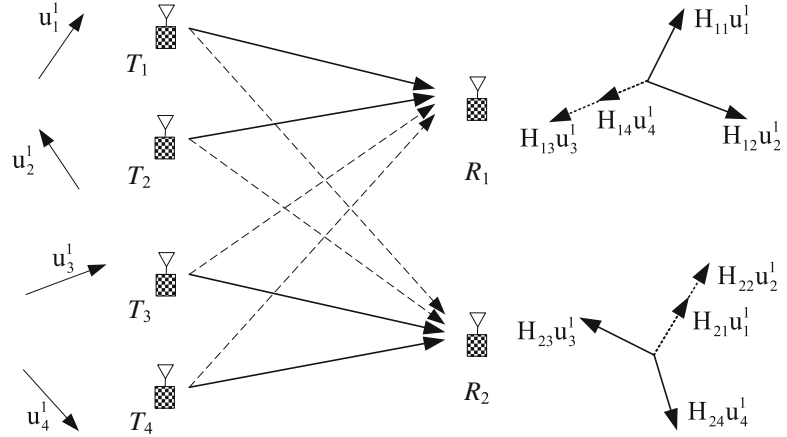
Interference Alignment in Wireless Networks,

Fig. 1 An example of SD-IA in a three-link wireless network. Solid lines represent data transmission, and dashed lines represent interference



Interference Alignment in Wireless Networks,

Fig. 2 An example of FD-IA. A solid arrow line represents data transmission, and a dashed arrow line represents interference



mission. Without FD-IA, this network can transport three data streams from the transmitters to the receivers (one on each subcarrier). By using FD-IA, this network can transport four data streams from the transmitters to the receivers (one on each link). The FD-IA scheme for this network is illustrated in Fig. 2. Denote $\mathbf{H}_{ji}$ as the frequency-domain channel matrix between receiver R_j and transmitter T_i . Due to the orthogonality of the frequency subcarriers, $\mathbf{H}_{ji}$ is a full-rank diagonal matrix, and its k th diagonal entry is channel coefficient of the k th subcarrier. Denote $\mathbf{u}_i^k$ as the precoding vector for the k th outgoing stream at transmitter T_i . For each receiver, it has two desired data streams and two interfering streams. Since the receiver has only three frequency subcarriers, it can decode both its desired data streams free of interference only if its two interfering streams are aligned in the same direction. One precoding scheme that can achieve the desired alignment at both receivers is as follows: $\mathbf{u}_1^1 = \mathbf{u}_{\text{ref}}$, $\mathbf{u}_2^1 = \mathbf{H}_{22}^{-1}\mathbf{H}_{21}\mathbf{u}_1$, $\mathbf{u}_3^1 = \mathbf{u}_{\text{ref}}$, $\mathbf{u}_4^1 = \mathbf{H}_{24}^{-1}\mathbf{H}_{23}\mathbf{u}_3$, where $\mathbf{u}_{\text{ref}}$ is a 3×1 reference vector with nonzero entries. By using the above precoding vectors, it can be verified that at each receiver, the two interfering streams will be aligned in the same direction, and the two desired data streams will lie in two independent directions, as shown in the figure. Therefore, four data streams can be transported over three frequency subcarriers in this network.

FD-IA in the above example appears similar to SD-IA in Fig. 1. But there is a fundamental difference between them. The design of FD-IA

relies on the characteristics of frequency-domain channels, which are diagonal matrices with each diagonal entry being the channel coefficient of a frequency subcarrier. In contrast, the design of SD-IA relies on the characteristics of space-domain channels, which are full (rather than diagonal) matrices with each entry being the channel coefficient from a transmit antenna to a receive antenna. Due to this fundamental difference, a direct extension of an IA scheme from the space domain to the frequency domain by simply treating an antenna element as a frequency subcarrier is not plausible and may result in an infeasible solution.

Time-Domain IA TD-IA refers to a joint construction of the signals at the transmitters in the time domain so that at each receiver, the interfering signals are aligned to the same direction, while the desired signals are free of interference. TD-IA can be achieved in different ways. For example, it can be achieved as the FD-IA scheme in Fig. 2 by simply treating each subcarrier as a time slot; it can also be achieved by leveraging the large propagation delays of signal travel (e.g., in underwater acoustic networks). In Grokop et al. (2011), Grokop et al. proposed a TD-IA scheme by exploiting the propagation delays between transmitters and receivers. They showed that by using TD-IA, the spectral efficiency of the K -user interference channel can grow linearly with K if the channel bandwidth is sufficiently large. In Chitre et al. (2012), Chitre et al. developed scheduling algorithms to exploit the benefits of

Interference Alignment in Wireless Networks, Table 1 A summary of SD-IA, FD-IA, and TD-IA

	Application scenarios	CSI at TX	Coordination among TX	SNR requirement at RX	Channel rank requirement
SD-IA	Multi-antenna networks	Required	Required	High SNR regime	Full-rank and independent channels
TD-IA	Networks with large propagation delays	Required	Required	All SNR regime	No requirement
FD-IA	OFDM networks	Required	Required	High SNR regime	Full-rank and frequency-selective channels

TD-IA for throughput improvement in a multiuser network with large propagation delays. In Zeng et al. (2017a), Zeng et al. showed that propagation-delay-based TD-IA can be achieved in a distributed fashion to improve the multi-hop network throughput.

Summary Table 1 summarizes the properties and requirements of these three IA schemes, including their application scenarios, channel state information (CSI) requirement, coordination requirement, signal-to-noise ratio (SNR) requirement, and channel rank requirement.

Key Applications

IA can be used to manage interference for wireless networks that have coordination among their transmitters. An application scenario of IA is enterprise Wi-Fi networks, where the access points (APs) have multiple antennas and are connected via high-speed cables. Another application scenario of IA is cellular wireless networks, where the base stations can cooperate with each other for interference management. In such networks, IA can be used to manage the co-channel (inter-cell) interference so as to improve the network throughput.

Cross-References

- [Interference Alignment for Power Line Communications](#)
- [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)

- [Interference-Aware Distributed Resource Allocation](#)
- [Interference-Aware Scheduling in Wireless Networks](#)

References

- Bresler G, Cartwright D, Tse D (2014) Feasibility of interference alignment for the MIMO interference channel. *IEEE Trans Inf Theory* 60(9): 5573–5586
- Cadambe VR, Jafar SA (2008) Interference alignment and degrees of freedom of the K -user interference channel. *IEEE Trans Inf Theory* 54(8):3425–3441
- Chitre M, Motani M, Shahabudeen S (2012) Throughput of networks with large propagation delays. *IEEE J Oceanic Eng* 37(4):645–658
- El Ayach O, Peters SW, Heath RW (2010) The feasibility of interference alignment over measured MIMO-OFDM channels. *IEEE Trans Veh Technol* 59(9): 4309–4321
- Gollakota S, Perli SD, Katabi D (2009) Interference alignment and cancellation. In: *ACM SIGCOMM computer communication review*, vol 39. ACM, pp 159–170
- Grokop LH, David N, Yates RD (2011) Interference alignment for line-of-sight channels. *IEEE Trans Inf Theory* 57(9):5820–5839
- Jafar SA (2011) Interference alignment – A new look at signal dimensions in a communication network. *Found Trends Commun Inf Theory* 7(1):1–134
- Lee N, Lim JB, Chun J (2010) Degrees of freedom of the MIMO Y channel: signal space alignment for network coding. *IEEE Trans Inf Theory* 56(7): 3332–3342
- Suh C, Tse D (2008) Interference alignment for cellular networks. In: 2008 46th annual Allerton conference on communication, control, and computing. IEEE, pp 1037–1044
- Suh C, Ho M, Tse DN (2011) Downlink interference alignment. *IEEE Trans Commun* 59(9):2616–2626

- Zeng H, Hou YT, Shi Y, Lou W, Kompella S, Midkiff SF (2017a) A distributed scheduling algorithm for underwater acoustic networks with large propagation delays. *IEEE Trans Commun* 65(3):1131–1145
- Zeng H, Shi Y, Hou YT, Lou W, Yuan X, Zhu R, Cao J (2017b) OFDM-based interference alignment in single-antenna cellular wireless networks. *IEEE Trans Commun* 65(10):4492–4506

Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks

Yong Zhou¹ and Weihua Zhuang²

¹School of Information Science and Technology, ShanghaiTech University, Shanghai, China

²Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

Synonyms

Interference correlation; Joint Laplace transform of co-channel interference; Poisson point process

Definitions

Interference characterization refers to mathematically characterizing the co-channel interference generated by concurrent transmissions, in terms of its probability density function, Laplace transform, or expectation, via modeling the influential sources of randomness, including the spatial locations of the interferers, the traffic pattern, the medium access scheme, and the propagation channel.

Background

Due to the scarcity of the radio spectrum, resource (e.g., time, frequency, and code) sharing among multiple wireless transceivers is necessary to enhance the radio spectral efficiency

of wireless ad hoc networks. Because of the broadcast nature of wireless medium, spatially separated concurrent transmissions in the same radio channel inevitably lead to co-channel interference, which is the main performance-limiting factor. As a result, the fundamental step toward analyzing the network-wide performance is the accurate characterization of co-channel interference due to concurrent transmissions.

Two widely used interference models are the protocol interference model and physical interference model (Cardieri 2010). Under the protocol interference model, the impact of co-channel interference on the packet reception is binary (Zhou and Zhuang 2013). In particular, the interference from an unintended transmitter located outside the interference range of a receiver has negligible impact on the packet reception. The interference from an unintended transmitter located within the interference range of a receiver always interrupts the packet reception. The interference range is set based on the signal-to-noise ratio (SNR) threshold required for successful packet reception. On the other hand, under the physical interference model, the power levels of the signals from all unintended transmitters are added, and the sum is considered as interference power at a receiver. A packet is successfully decoded if the signal-to-interference-plus-noise ratio (SINR) observed at the receiver is not smaller than the required reception threshold.

Interference characterization has been attracting more and more attention in recent years. In Gilhousen et al. (1991), the co-channel interference was approximated by a Gaussian random variable, where the mean and variance of co-channel interference are set by using an empirical fitting method. The Gaussian approximation enables tractable performance analysis for fixed topology networks. The Wyner model (Wyner 1994) is another simplified model proposed for interference characterization, which only takes into account the nearest neighboring interferers. Both the Gaussian approximation and the Wyner model cannot accurately characterize the co-channel interference in wireless ad hoc networks, as the co-channel interference depends

on multiple sources of randomness, including the interferers' spatial distribution, the traffic pattern, the medium access scheme, and the propagation channel. Besides, conducting exhaustive simulations can overcome the shortcomings of the Gaussian approximation and the Wyner model by averaging out all the randomness. However, the simulation-based method is always time-consuming and error-prone.

In order to incorporate the randomness of the interferers' spatial locations into interference characterization, a more advanced model based on the point process theory was proposed in Baccelli and Błaszczyszyn (2010). In particular, the spatial locations of concurrent transmitters are modeled as a homogeneous Poisson point process (PPP). By using tools from stochastic geometry, the co-channel interference at any time instant can be characterized in terms of its Laplace transform. Such a model enables accurate performance evaluation for both wireless ad hoc networks (Baccelli et al. 2009) and cellular networks (ElSawy et al. 2017). Based on the Poisson shot noise field theory, the authors in Baccelli et al. (2009) analyzed the outage probability of an opportunistic ALOHA scheme in a mobile ad hoc network (MANET), where the randomness of both the medium access and the spatial locations of the interferers is considered. Moreover, the accuracy of the PPP modeling for the spatial locations of the base stations in large-scale cellular networks was demonstrated in ElSawy et al. (2017).

Cooperative communication is an effective technique, which has the potential to mitigate the detrimental effect of wireless channel impairments by achieving a spatial diversity gain (Sendonaris et al. 2003; Zhuang and Zhou 2013). The process of cooperative communication is generally composed of two phases, namely, information sharing and cooperative transmission phases. In the information sharing phase, the transmitter broadcasts a packet to the potential relays and the intended receiver. In the cooperative transmission phase, the potential relays that overheard the packet act as relays and help forward the packet to the intended receiver. By combining two or more copies of the same

packet transmitted through independent links, a spatial diversity gain is achieved at the receiver. The advantages of cooperative communication in the physical layer have been demonstrated by analyzing different relaying strategies from the information theory point of view. For different application scenarios, the cooperative advantages at the physical layer can be transformed into specific advantages at the link layer, such as enhancing transmission reliability, reducing transmission power, increasing transmission rate, and enlarging network coverage (Shan et al. 2009; Zhuang and Ismail 2012).

Interference Characterization for Cooperative Communication

Different from a direct link, a cooperative link is characterized by the extra relay(s) and two-phase transmission. As a result, the co-channel interference generated by a cooperative link is different from the co-channel interference generated by a direct link, as an additional relay or relays participate in the transmission of one packet. In other words, the cooperative links redistribute the interference over a large-scale wireless network due to relay transmissions (Zhou and Zhuang 2015, 2016). In order to accurately characterize the network-wide performance of cooperative communication, the interference redistribution due to concurrent cooperative transmissions should be taken into account. Thereby, an interference characterization framework for cooperative communication under the physical interference model is presented as follows.

Consider a two-dimensional network coverage area with spatially random transmitters, relays, and receivers. The spatial locations of the transmitters are assumed to follow a homogeneous PPP, denoted as $\Phi_S = \{s_1, s_2, \dots\}$, with density λ_S , which represents the average number of transmitters per unit area. Each transmitter (e.g., S_i) is associated with a receiver (e.g., D_i), which is located at L meters away with a random direction. The relays (e.g., R_i) do not have their own packets to transmit and receive, and they are always willing to help forward packets for

the transmitters. The spatial locations of the relays also form a homogeneous PPP, denoted as $\Phi_R = \{r_1, r_2, \dots\}$, with density λ_R . Due to the stationary property of the homogeneous PPP, the co-channel interference observed at the typical transmitter-receiver pair, indexed as $i = 0$, follows the same distribution with that at any transmitter-receiver pair in the network. Without loss of generality, the typical transmitter and receiver are located at $(L, 0)$ and $(0, 0)$, respectively. The channel between any two transceivers suffers from Rayleigh fading and path loss. All channel fading coefficients are assumed to be independent and identically distributed (i.i.d.) random variables with unit mean. The path loss between transmitter S_i and receiver D_j is denoted as $g(s_i - d_j)$.

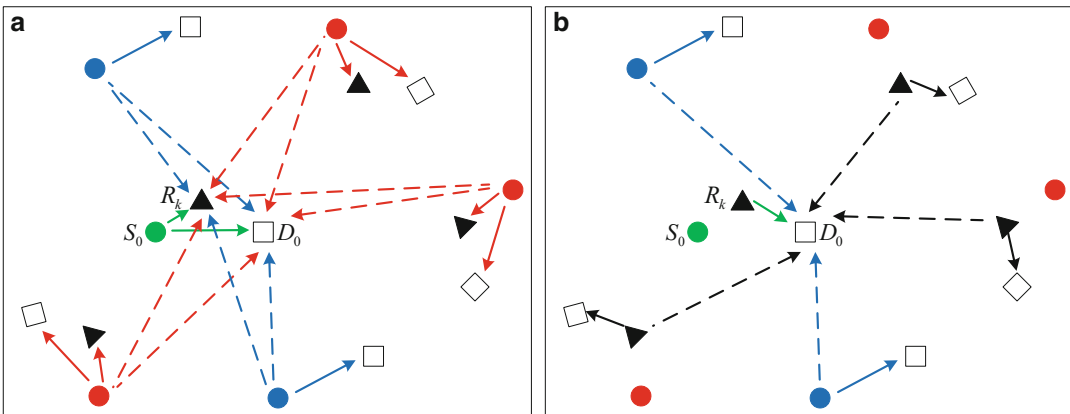
All transmitters and relays transmit with the same power, which is normalized to one. Each transmitter always has a packet for transmission and makes an independent decision on the transmission mode (i.e., direct transmission or cooperative transmission). Different criteria (e.g., the number of relays available) can be applied to enable opportunistic cooperation. This entry considers a general case and denotes the probability that each transmitter-receiver pair enables cooperative transmission as p . Hence, by applying the independent thinning property of the PPP, the

final cooperative transmission set (consisting of the spatial locations of transmitters of cooperative links) and direct transmission set are independent and denoted as Φ_C and Φ_D with densities $\lambda_C = p\lambda_S$ and $\lambda_D = (1 - p)\lambda_S$, respectively.

Relay selection plays a pivotal role in determining the performance of a cooperative transmission scheme. An efficient relay selection scheme should select the best relay among randomly positioned relays while taking into account the spatial frequency reuse. Various relay selection schemes (e.g., Zhou and Zhuang 2017) can be applied. This entry does not specify the relay selection scheme, and the interference characterization method presented here can be modified to fit specific relay selection schemes.

For both direct and cooperative transmission, each packet is transmitted in one time slot. For cooperative transmission, each time slot is divided equally into two sub-time slots. In the first sub-time slot, each transmitter transmits a packet to its intended receiver. As shown in Fig. 1a, the aggregate interference power observed at the relay (e.g., R_k) of the typical transmitter-receiver pair in the first sub-time slot can be expressed as

$$I_{R_k:1}(\Phi_D, \Phi_C) = I_{DR_k:1}(\Phi_D) + I_{CR_k:1}(\Phi_C), \quad (1)$$



Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks, Fig. 1 An illustration of the interference power, originating from both the direct and cooperative links over the network, observed by relay R_k and destination D_0 in the first and second sub-time slots. Each circle, triangle, and square

represent a source, selected relay, and destination, respectively. The solid and dashed lines represent the transmitted signal and interference, respectively. (Reproduced with permission from Springer.) (a) Transmission in the first sub-time-slot. (b) Transmission in the second sub-time-slot

where $I_{DR_k:1}(\Phi_D) = \sum_{s_i \in \Phi_D} H_{S_i R_k:1} g(s_i - r_k)$ is the interference power at relay R_k from the transmitters of direct links, $I_{CR_k:1}(\Phi_C) = \sum_{s_j \in \Phi_C} H_{S_j R_k:1} g(s_j - r_k)$ denotes the interference power at relay R_k from the transmitters of cooperative links, and $H_{S_i R_k:1}$ denotes the fading coefficient between transmitter S_i and relay R_i in the first sub-time slot.

Taking into account the interference redistribution incurred by cooperative transmission, as shown in Fig. 1b, the aggregate interference power observed at receiver D_0 in the first and second sub-time slots can be expressed respectively as

$$\begin{aligned} I_{D_0:1}(\Phi_D, \Phi_C) &= I_{DD_0:1}(\Phi_D) + I_{CD_0:1}(\Phi_C), \\ I_{D_0:2}(\Phi_D, \Phi_F) &= I_{DD_0:2}(\Phi_D) + I_{FD_0:2}(\Phi_F), \end{aligned} \quad (2)$$

where $I_{DD_0:1}(\Phi_D) = \sum_{s_i \in \Phi_D} H_{S_i D_0:1} g(s_i)$ and $I_{DD_0:2}(\Phi_D) = \sum_{s_i \in \Phi_D} H_{S_i D_0:2} g(s_i)$ denote the aggregate interference powers from the transmitters of direct links in the first and second sub-time-slots, respectively, and $I_{CD_0:1}(\Phi_C) = \sum_{s_j \in \Phi_C} H_{S_j D_0:1} g(s_j)$ and $I_{FD_0:2}(\Phi_F) = \sum_{r_m \in \Phi_F} H_{R_m D_0:2} g(r_m)$ denote the aggregate interference powers from the

transmitters and selected relays of cooperative links in the first and second sub-time slots, respectively. Here, Φ_F denotes the PPP formed by the spatial locations of the selected relays that forward the packets in the second sub-time slot.

As the spatial locations of the transmitters of direct links in both sub-time slots are the same, interference powers $I_{DD_0:1}(\Phi_D)$ and $I_{DD_0:2}(\Phi_D)$ are temporally correlated. Because the transmitter (e.g., S_j) is always located adjacent to its selected relay (e.g., R_m) for each cooperative link, interference powers $I_{CD_0:1}(\Phi_C)$ and $I_{FD_0:2}(\Phi_F)$ are also temporally correlated. In summary, the interference powers observed at receiver D_0 in two sub-time slots, i.e., $I_{D_0:1}(\Phi_D, \Phi_C)$ and $I_{D_0:2}(\Phi_D, \Phi_F)$, are temporally correlated. On the other hand, the relay of the typical transmitter-receiver pair and the typical receiver are geographically close, and they suffer from the interference generated by the same or adjacent interferers. Hence, the interference powers $I_{R_k:1}(\Phi_D, \Phi_C)$ and $I_{D_0:1}(\Phi_D, \Phi_C)$ or $I_{D_0:2}(\Phi_D, \Phi_F)$ are spatially correlated.

Based on the above discussions, the joint Laplace transform of interference powers $I_{D_0:1}(\Phi_D, \Phi_C)$ and $I_{D_0:2}(\Phi_D, \Phi_F)$ is given by

$$\begin{aligned} \mathcal{L}(\zeta_1, \zeta_2) &= \mathbb{E} \left[\exp(-\zeta_1 I_{D_0:1}(\Phi_D, \Phi_C) - \zeta_2 I_{D_0:2}(\Phi_D, \Phi_F)) \right] \\ &\stackrel{(a)}{=} \mathbb{E} \left[\exp(-\zeta_1 I_{DD_0:1}(\Phi_D) - \zeta_2 I_{DD_0:2}(\Phi_D)) \right] \\ &\quad \times \mathbb{E} \left[\exp(-\zeta_1 I_{CD_0:1}(\Phi_C) - \zeta_2 I_{FD_0:2}(\Phi_F)) \right], \end{aligned} \quad (3)$$

where (a) follows by substituting (2) and from the independence between PPP Φ_D and PPP Φ_C or PPP Φ_F .

By taking the Laplace transform of the fading coefficients between the interferers and the typical receiver, $\mathcal{A}_1 = \mathbb{E} \left[\exp(-\zeta_1 I_{DD_0:1}(\Phi_D) - \zeta_2 I_{DD_0:2}(\Phi_D)) \right]$ can be computed by

$$\begin{aligned} \mathcal{A}_1 &= \mathbb{E} \left[\prod_{s_i \in \Phi_D} \frac{1}{(1 + \zeta_1 g(s_i)) (1 + \zeta_2 g(s_i))} \right] \\ &\stackrel{(a)}{=} \exp \left(-\lambda_D \int_{\mathbb{R}^2} \left(1 - \frac{1}{(1 + \zeta_1 g(s)) (1 + \zeta_2 g(s))} \right) ds \right), \end{aligned} \quad (4)$$

where (a) follows from the probability generating functional (PGFL) of the PPP.

Similarly, $\mathcal{A}_2 = \mathbb{E}[\exp(-\zeta_1 I_{CD_0:1}(\Phi_C) - \zeta_2 I_{FD_0:2}(\Phi_F))]$ is given by

$$\begin{aligned} \mathcal{A}_2 &= \mathbb{E} \left[\left(\prod_{s_i \in \Phi_C} \frac{1}{1 + \zeta_1 g(s_i)} \right) \left(\prod_{r_m \in \Phi_F} \frac{1}{1 + \zeta_2 g(r_m)} \right) \right] \\ &\stackrel{(a)}{=} \mathbb{E} \left[\left(\prod_{s_i \in \Phi_C} \frac{1}{1 + \zeta_1 g(s_i)} \right) \left(\prod_{s_i \in \Phi_C} \left(\frac{1 - q_e}{1 + \zeta_2 g(s_i + \tau)} + q_e \right) \right) \right] \\ &\stackrel{(b)}{=} \exp \left(-\lambda_C \int_{\mathbb{R}^2} \left(1 - \frac{1}{1 + \zeta_1 g(s)} \left(\frac{1 - q_e}{1 + \zeta_2 g(s + \mathbb{E}_\tau[\tau])} + q_e \right) \right) ds \right), \end{aligned} \quad (5)$$

where q_e is probability that no relays are available, τ denotes the coordinate difference between the transmitter and the selected relay, (a) follows from the transformation between PPP Φ_C and PPP Φ_F , and (b) follows from the PGFL of the PPP. Note that both q_e and τ depend on the adopted relay selection scheme.

By substituting (4) and (5) into (3), the joint Laplace transform of interference powers

$I_{D_0:1}(\Phi_D, \Phi_C)$ and $I_{D_0:2}(\Phi_D, \Phi_F)$ can be obtained. Following the similar steps, the Laplace transform of each interference power (e.g., $I_{D_0:1}(\Phi_D, \Phi_C)$ or $I_{D_0:2}(\Phi_D, \Phi_F)$) can also be derived.

On the other hand, the joint expectation of interference powers $I_{D_0:1}(\Phi_D, \Phi_C)$ and $I_{D_0:2}(\Phi_D, \Phi_F)$ is another important metric, which is given by

$$\begin{aligned} \mathcal{E} &= \mathbb{E} [I_{D_0:1}(\Phi_D, \Phi_C) I_{D_0:2}(\Phi_D, \Phi_F)] \\ &= \mathbb{E} [I_{DD_0:1}(\Phi_D) I_{DD_0:2}(\Phi_D)] + \mathbb{E} [I_{DD_0:1}(\Phi_D) I_{FD_0:2}(\Phi_F)] \\ &\quad + \mathbb{E} [I_{CD_0:1}(\Phi_C) I_{DD_0:2}(\Phi_D)] + \mathbb{E} [I_{CD_0:1}(\Phi_C) I_{FD_0:2}(\Phi_F)]. \end{aligned} \quad (6)$$

As PPP Φ_D is independent of PPP Φ_C and PPP Φ_F , the following equations hold

$$\mathbb{E} [I_{CD_0:1}(\Phi_C) I_{DD_0:2}(\Phi_D)] = \lambda_C \lambda_D \left(\int_{\mathbb{R}^2} g(s) ds \right)^2. \quad (7)$$

$$\mathbb{E} [I_{DD_0:1}(\Phi_D) I_{FD_0:2}(\Phi_F)] = \lambda_D \lambda_F \left(\int_{\mathbb{R}^2} g(s) ds \right)^2,$$

As PPP Φ_C and PPP Φ_F are not independent of each other, the mean product of interference powers $I_{CD_0:1}(\Phi_C)$ and $I_{FD_0:2}(\Phi_F)$ is given by

$$\begin{aligned} \mathbb{E} [I_{CD_0:1}(\Phi_C) I_{FD_0:2}(\Phi_F)] &\stackrel{(a)}{=} \mathbb{E} \left[\left(\sum_{s_j \in \Phi_C} g(s_j) \right) \left(\sum_{r_m \in \Phi_F} g(r_m) \right) \right] \\ &\stackrel{(b)}{=} \mathbb{E} \left[\left(\sum_{s_j \in \Phi_C} g(s_j) \right) \left(\sum_{s_j \in \Phi_C} (1 - q_e) g(s_j + \tau) \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left[\sum_{s_j \in \Phi_C} (1 - q_e) g(s_j) g(s_j + \tau) \right] \\
&\quad + \mathbb{E} \left[\sum_{s_i, s_j \in \Phi_C, s_i \neq s_j} (1 - q_e) g(s_j) g(s_j + \tau) \right] \\
&\stackrel{(c)}{=} \lambda_F \int_{\mathbb{R}^2} g(s) \mathbb{E}_\tau [g(s + \tau)] ds + \lambda_C \lambda_F \left(\int_{\mathbb{R}^2} g(s) ds \right)^2, \tag{8}
\end{aligned}$$

where $\lambda_F = \lambda_C(1 - q_e)$ denotes the spatial density of PPP Φ_F , (a) follows as the channel fading coefficients are independent random variables with unit mean, (b) follows from the transformation between PPP Φ_C and PPP Φ_F , and (c) follows from Campbell's theorem and second-order product density formula of the PPP.

Similarly, by setting $\tau = (0, 0)$, the following equation holds

$$\begin{aligned}
&\mathbb{E} [I_{D_{0:1}}(\Phi_D) I_{D_{0:2}}(\Phi_D)] \\
&= \lambda_D \int_{\mathbb{R}^2} (g(s))^2 ds + \lambda_D^2 \left(\int_{\mathbb{R}^2} g(s) ds \right)^2. \tag{9}
\end{aligned}$$

By substituting (7), (8), and (9) into (6), the joint expectation of interference powers $I_{D_{0:1}}(\Phi_D, \Phi_C)$ and $I_{D_{0:2}}(\Phi_D, \Phi_F)$ can be obtained. Following the similar steps, the expectation of each interference power (e.g., $I_{D_{0:1}}(\Phi_D, \Phi_C)$ or $I_{D_{0:2}}(\Phi_D, \Phi_F)$) can also be obtained.

The joint Laplace transform and joint expectation of co-channel interference due to concurrent cooperative transmissions are useful for performance analysis in various large-scale networks. For example, they can be used to evaluate the spatial and temporal correlation levels of co-channel interference, to derive the network-wide performance (e.g., outage probability, achievable ergodic rate, and transmission capacity), and to estimate the average energy that can be harvested by exploiting co-channel interference.

Conclusions

In this entry, the backgrounds of the interference model, the interference characterization, and the cooperative communication are reviewed. Then, an interference characterization framework for cooperative communication is presented, where the co-channel interference is characterized in terms of the joint Laplace transform and the joint expectation. Finally, the applications of interference characterization are discussed.

Cross-References

- [Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space](#)
- [Path Loss and Interference Analysis for a Random Two-Hop Relay Link](#)

References

- Baccelli F, Błaszczyszyn B (2010) Stochastic geometry and wireless networks: volume II applications. Found Trends® Netw 4(2):1–312
- Baccelli F, Błaszczyszyn B, Muhlethaler P (2009) Stochastic analysis of spatial and opportunistic ALOHA. IEEE J Sel Areas Commun 27(7):1105–1119
- Cardieri P (2010) Modeling interference in wireless ad hoc networks. IEEE Commun Surv Tutorials 12(4):551–572
- ElSawy H, Sultan-Salem A, Alouini MS, Win MZ (2017) Modeling and analysis of cellular networks using stochastic geometry: a tutorial. IEEE Commun Surv Tutorials 19(1):167–203

- Gilhausen KS, Jacobs IM, Padovani R, Viterbi AJ, Weaver LA, Wheatley CE (1991) On the capacity of a cellular CDMA system. *IEEE Trans Veh Technol* 40(2): 303–312
- Sendonaris A, Erkip E, Aazhang B (2003) User cooperation diversity. Part I. System description. *IEEE Trans commun* 51(11):1927–1938
- Shan H, Zhuang W, Wang Z (2009) Distributed cooperative MAC for multihop wireless networks. *IEEE Commun Mag* 47(2):126–133
- Wyner AD (1994) Shannon-theoretic approach to a Gaussian cellular multiple-access channel. *IEEE Trans Inf Theory* 40(6):1713–1727
- Zhou Y, Zhuang W (2013) Beneficial cooperation ratio in multi-hop wireless ad hoc networks. In: *Proceeding of IEEE Infocom*, Turin, Apr 2013
- Zhou Y, Zhuang W (2015) Throughput analysis of cooperative communication in wireless ad hoc networks with frequency reuse. *IEEE Trans Wirel Commun* 14(1):205–218
- Zhou Y, Zhuang W (2016) Performance analysis of cooperative communication in decentralized wireless networks with unsaturated traffic. *IEEE Trans Wirel Commun* 15(5):3518–3530
- Zhou Y, Zhuang W (2017) Opportunistic cooperation in wireless ad hoc networks with interference correlation. *Peer-to-Peer Netw Appl* 10(1):238–252
- Zhuang W, Ismail M (2012) Cooperation in wireless communication networks. *IEEE Wirel Commun* 19(2): 10–20
- Zhuang W, Zhou Y (2013) A survey of cooperative MAC protocols for mobile communication networks. *J Internet Technol* 14(4):541–559

Interference Correlation

- [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)

Interference Management

- [Resource Allocation in NOMA-Assisted D2D Networks](#)

Interference Mitigation

- [Interference Alignment for Power Line Communications](#)

Interference Mitigation in Wireless Communications-Based Train Control (CBTC), Cognitive Control Approach

Hongwei Wang¹ and Richard Fei Yu^{2,3}

¹National Research Center of Railway Safety Assessment, Beijing Jiaotong University, Beijing, China

²Carleton University, Ottawa, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

[Communication-based Train Control \(CBTC\)](#); [Closed-circuit television \(CCTV\)](#); [Jamming](#); [Security](#)

Definitions

As a key component of urban rail transit systems, *communications-based train control* (CBTC) is an automated train control system using train-ground communications to ensure efficient operation of rail vehicles. In addition to CBTC systems, passenger information systems (PISs) are adopted in urban rail transit systems to improve quality of service (QoS) offered to customers. The interference between CBTC systems and PISs is an important factor impacting QoS of both CBTC systems and PISs. With recent advances in cognitive dynamic systems, a cognitive control approach is proposed to interference mitigation considering the co-existence of CBTC systems and PISs, where the notion of information gap is adopted to quantitatively describe effects of interference on CBTC and the wireless channel is modeled as a finite-state Markov chain with multiple state transition probability matrices derived from real-field measurements.

Background

Urban rail transit systems are developing rapidly around the world. Due to the huge urban traffic pressure, improving efficiency of urban rail transit systems is in high demand. *Communications-based train control* (CBTC) is adopted to improve utilization of railway network infrastructure and enhance the level of service offered to customers (Zhu et al. 2012). In addition to CBTC systems, *passenger information systems* (PISs) are adopted in urban rail transit systems to improve the quality of service (QoS) offered to customers (Zhu et al. 2011). PISs can provide real-time services (e.g., video) information to passengers at the station or onboard, including arriving and departure time, news, sport games, and advertisements. When emergencies (e.g., fire and terrorist attacks) happen, PISs can supply signs of emergency evacuation. Both CBTC systems and PISs use wireless local area networks (WLANs) as the main method for train-ground communications due to available commercial off-the-shelf (COTS) WLAN equipment with low costs. As a result, *interference* between CBTC systems and PISs is an important factor impacting QoS of both CBTC systems and PISs.

Real-field measurements on coexisting CBTC systems and PISs at Beijing Subway Line 1 show that percentage of packet losses is increased with interference from the PIS even if the associated AP does not change, besides those caused by handoffs (Wang et al. 2015), where CBTC APs working at channel 1 and PIS APs working at channel 6 and there is no overlapping frequency band.

Based on 802.11 protocols, a WLAN station applies the Clear Channel Assessment (CCA) mechanism to detect the channel state (busy or idle). According to measurement results, it is possible that the power of the PIS wireless link “leaks” into the adjacent channel (i.e., channel 1), which can cause the CBTC WLAN equipment to misjudge the channel state. Therefore, the interfered equipment could perform unnecessary backoff procedures or suffered from collisions,

which can lead to packet losses. As a result, cognitive control could be used to mitigate the inference issues (Wang et al. 2017).

A Cognitive Control Approach to Mitigate Impacts of Interference from the PIS

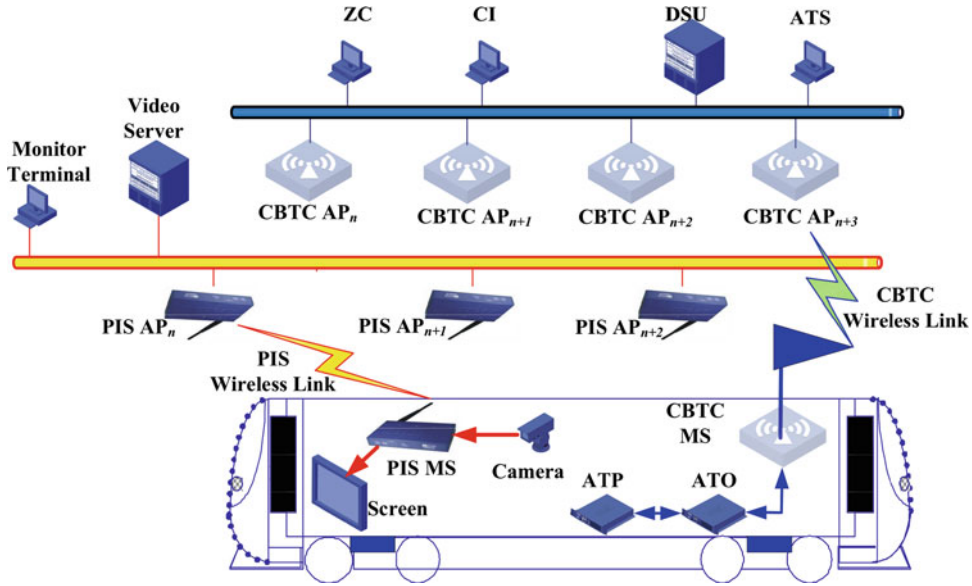
Continuous bidirectional wireless communications between the mobile station (MS) and access points (APs) are adopted in modern CBTC systems to guarantee high efficiency and reliability of train operation, which can connect wayside equipment and onboard equipment. Similarly to CBTC systems, a PIS consists of onboard devices, ground equipment, and a network. Onboard devices include cameras, screens, and PIS MSs, while ground equipment includes a video server, a monitor center, and PIS APs. PIS train-ground communications ensure passengers onboard can get real-time information from the video server and the monitor center can real-time receive monitoring pictures of the train interior and the station, as shown in Fig. 1.

The proposed cognitive control approach dynamically makes decisions, such as the packet size, the transmission interval, handoff decisions, and physical layer parameters, in order to guarantee the data transmission between the train and the control center in CBTC systems. Adding a new term about information gap compared with traditional cost function of a quadratic optimal controller, the resulting cost function can be formulated as follows (Haykin et al. 2012).

$$J = (s - \tilde{s})^T W (s - \tilde{s}) + u^T R u + \beta G \quad (1)$$

where G is the information gap and β is a scalar, $\tilde{s}$ is the desired state of the system, s is the actual state of the system, u is the physical control vector, and matrices W and R are applied as desired weights for systems's state and control.

As a result, a cognitive controller on the train should help the CBTC MS to make some decisions through reinforcement learning



Interference Mitigation in Wireless Communications-Based Train Control (CBTC), Cognitive Control Approach, Fig. 1 Coexistence of a CBTC system and a PIS

(RL) (Yu et al. 2006) in order to improve communication performance, which could be implemented at the application layer, the MAC layer, and the physical layer. The action determined at each communication cycle at the k th communication cycle could be denoted as $a_k = \{a_k^s, a_k^i, a_k^h, a_k^m\}$, where a_k^s is the packet size action, a_k^i is the packet interval action, a_k^h is the handoff action, and a_k^m is the multiplexing gain action. The state s_k is denoted as $s_k = \{\gamma_{1k}^l, \gamma_{2k}^l, ID, PLR, PI, PS\}$, where ID is the identification number of the current associated AP, PLR is the packet loss rate, PI is the packet interval, and PS is the packet size. When ID changes, the handoff procedure happens.

According to Eq. (1), the information gap can be defined based on the QoS. QoS indexes could be given as D_k, L_k , where D_k and L_k are the time delay and the number of packet loss at the k th communication cycle and $G_k = \{D_k, L_k\}$. Therefore, the cost function of cognitive control in the paper is shown as follows:

$$J = (s_k - \hat{s}_k) Q (s_k - \hat{s}_k)^T + a_k R a_k^T + \beta G_k G_k^T \quad (2)$$

where $\hat{s}_k$ is the desirable state of the system.

Formulation of the Cognitive Control Approach

The MAC layer model and the physical layer model are needed to formulate the cognitive control approach to mitigate impacts of interference from the PIS. According to the structure of the cognitive control system, the CBTC MS is considered as the perceptual part that can calculate the information gap according to QoS indexes. Based on the information gap, the cognitive controller can make decisions that should be performed on the CBTC MS.

According to the random access scheme of CSMA/CA protocols, a random packet is transmitted after the DIFS time and the backoff time, which is uniformly distributed in the range $[0, CW - 1]$. As a result, $BackoffTime_\alpha = Random([0, CW_\alpha - 1]) \times aSlotTime$, where $aSlotTime$ is a constant time corresponding to IEEE 802.11g standards. CW is the value of contention window which has lower and upper bounds denoted as CW_{min} and CW_{max} . For the first backoff, CW is initialized as CW_{min} . Then, each backoff will double CW until it reaches CW_{max} . The contention window

is kept as $CW_{\max}$ when times of backoff are no less than m . As a result, the probability

$BackoffTime = i$ could be obtained as follows.

$$P(BackoffTime = i) = \begin{cases} \sum_{\alpha=0}^{m-1} \frac{p^\alpha(1-p)}{2^\alpha CW_{\min}} + \frac{p^m}{CW_{\max}} & 1 \leq i \leq CW_{\min} \\ \sum_{\alpha=\lfloor \log_2 i \rfloor}^{m-1} \frac{p^\alpha(1-p)}{2^\alpha CW_{\min}} + \frac{p^m}{CW_{\max}} & 2^{\lfloor \log_2 i \rfloor - 1} CW_{\min} + 1 \leq i \leq 2^{\lfloor \log_2 i \rfloor} CW_{\min} \\ \frac{p^m}{CW_{\max}} & 2^{m-1} CW_{\min} - 1 \leq i \leq CW_{\max} \end{cases} \quad (3)$$

where $\lfloor \cdot \rfloor$ is a function to round numbers, p is the probability of unsuccessful transmissions caused by the collision of the MAC layer and the channel fading of the physical layer, and α is times of backoff. Therefore, the expected value of backoff time is as follows.

$$\begin{aligned} \overline{BackoffTime} &= (1-p) \frac{CW_{\min}}{2} \\ &+ \dots + p^m(1-p) \frac{\sum_{i=0}^m 2^i CW_{\min}}{2} + \\ &\dots + p^{m_r} \frac{\sum_{i=0}^m 2^i CW_{\min} + (m_r - m) 2^m CW_{\min}}{2} \end{aligned} \quad (4)$$

For the periodic PIS service, the uplink is the CCTV, while the downlink is the PIS video stream. Packet sizes of PIS and CCTV are 20480 Bytes and 10240 Bytes, respectively. According to 802.11 MAC protocols, the maximum MPDU (MAC Protocol Data Unit) length is 4096 Bytes. As a result, the PIS frame and CCTV frame should be fragmented, and the maximum fragmentation of a MPDU is 3000 Bytes. A PIS frame could be divided into $\frac{VideoFrameLength}{3000 \times 8}$ fragments. According to the fragmented MPDU transmission procedure, the duration for a successful transmission of a PIS frame is calculated as follows:

$$T_{\text{video}} = \frac{L_v}{R} + (aSIFSTime + anACKTime) \times \lceil \frac{L_v}{l_m} \rceil \quad (5)$$

where R is the transmission rate, $anACKTime$ is the time for the transmitter to send an acknowledge frame, $\lceil \cdot \rceil$ is the top integral function, L_v is the length of a video frame, and l_m is the maximum length of MPDU stated in 802.11 protocols.

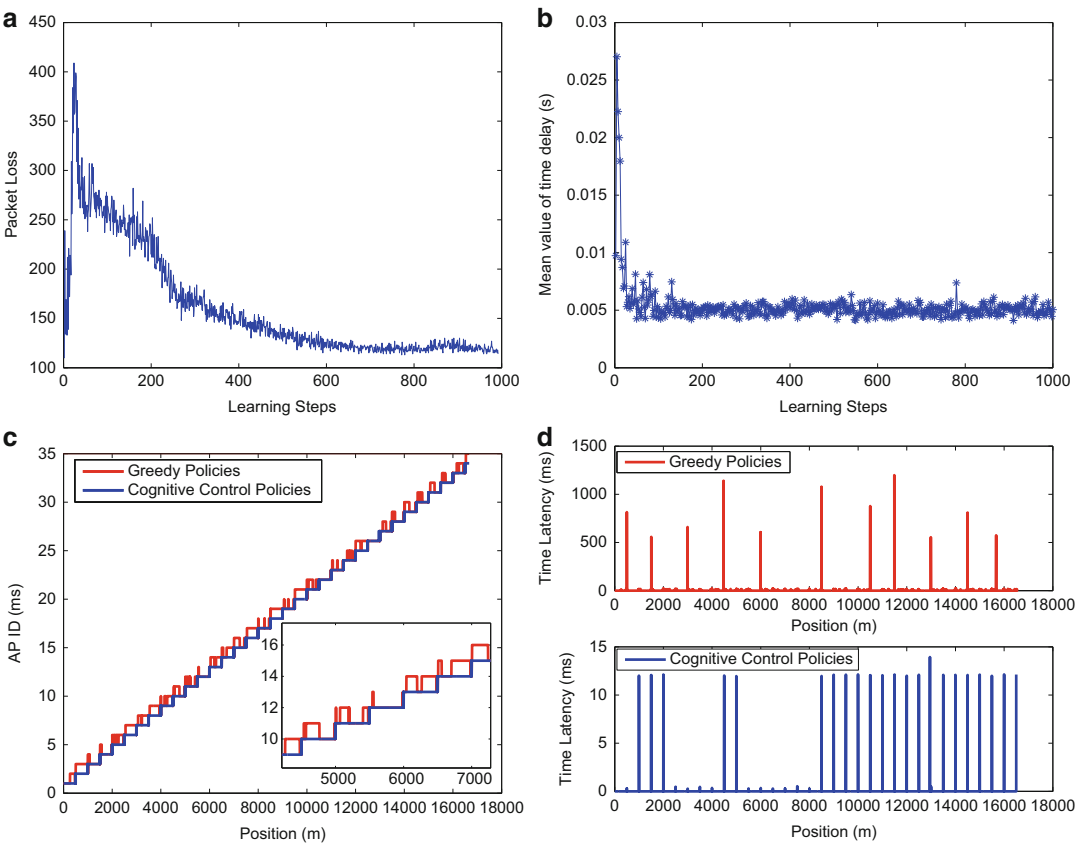
Based on the multi-matrices FSMC model in Wang et al. (2014), strength of interference signals could be obtained. As the receiver is equipped with a band-width filter, we could apply the filter's frequency response to calculate the leaked interference signal. Assume the filter's frequency response function is $F(f)$ and the power spectral density is $P(f)$. According to the channel spacing of 802.11g, the interference factor

can be derived as $IF = \frac{\int_{-\frac{w_a}{2}}^{\frac{w_a}{2}} P(f)F(f)df}{\int_{-\frac{w}{2}}^{\frac{w}{2}} P(f)df}$, where

w_a is the bandwidth of the adjacent channel and w is the bandwidth of the interference signal. Therefore, IF can be used to calculate the interference signal leaked into the adjacent channel based on the multi-matrices FSMC model.

Simulation Results and Discussions

Matlab is used to implement simulations, where there are 35 APs and the coverage of each AP is 500 m according to Wang et al. (2014). Figure 2a shows that the Q learning coverages about 600 iterations and packet loss rate decreases with the increase of learning times, which means optimal policies have been obtained. With the learning process being implemented, mean value of time latency is becoming lower and lower as shown in Fig. 2b, where it is about 5 ms when cognitive control converges to a stable state. Fig-



Interference Mitigation in Wireless Communications-Based Train Control (CBTC), Cognitive Control Approach, Fig. 2 Simulation results. (a) Packet losses versus learning steps. (b) Mean value of time latency

between consecutive instants of handoffs. (c) Handoff decisions under different policies. (d) Handoff latency under different policies

ure 2c demonstrates that cognitive control policies can almost eliminate the ping-pong handoffs. Besides, Fig. 2d shows that handoff latency under cognitive control policies is dozens of milliseconds, while handoff latency under traditional greedy policies is much higher close to 1 s.

the cognitive control approach can significantly reduce the packet losses, time delay, and handoff latency under interference from the coexisting PIS.

Conclusions and Future Work

Cognitive control is used to improve performance of CBTC train-ground communications under interference from co-existing PISs. Based on the cognitive control formulation, reinforcement learning was used to get the optimal strategy. Simulation results were presented to show that

Key Applications

The interference mitigation in wireless CBTC systems has been designed to increase the availability and reliability of train operation. The cognitive control based approach can real-timely make the interference mitigation decisions implemented on the physical layer, MAC layer and application layer.

References

- Haykin S, Fatemi M, Setoodeh P, Xue Y (2012) Cognitive control. *Proc IEEE* 100(12):3156–3169
- Wang H, Yu FR, Zhu L, Tang T, Ning B (2014) Finite-state Markov modeling for wireless channels in tunnel communication-based train control systems. *IEEE Trans Intell Transp Syst* 15(3):1083–1090
- Wang H, Yu FR, Zhu L, Tang T, Ning B (2015) A cognitive control approach to communication-based train control systems. *IEEE Trans Intell Transp Syst* 16(4):1676–1689
- Wang H, Yu FR, Wang H (2017) A cognitive control approach to interference mitigation in communications-based train control (CBTC) co-existing with passenger information systems (PISs). *EURASIP J Wirel Commun Netw* 2017(1):186
- Yu F, Wong VW, Leung V (2006) Efficient QoS provisioning for adaptive multimedia in mobile communication networks by reinforcement learning. *ACM/Springer Mobile Netw Appl J (MONET)* 11(1):101–110
- Zhu L, Yu FR, Ning B, Tang T (2011) Cross-layer design for video transmissions in metro passenger information systems. *IEEE Trans Veh Technol* 60(3):1171–1181
- Zhu L, Yu FR, Ning B, Tang T (2012) Cross-layer handoff design in MIMO-enabled WLANs for communication-based train control (CBTC) systems. *IEEE J Select Areas Commun* 30(4):719–728

Interference-Aware Distributed Resource Allocation

Hai Jiang
Department of Electrical and Computer
Engineering, University of Alberta, Edmonton,
AB, Canada

Synonyms

[Resource management](#); [Medium access control](#)

Definitions

Interference-aware distributed resource allocation refers to a distributed medium access scheme supporting multiple transmission links in a wireless ad hoc network, in which each

transmission link is aware of the interference level that it receives from other links.

Historical Background

A wireless ad hoc network is distributed in nature, and thus, a distributed medium access control (MAC) protocol is preferred. In the literature, the most popular distributed MAC protocols are based on carrier sense multiple access (CSMA) (Kleinrock and Tobagi 1975) or busy tone (Haas and Deng 2002; Wang and Zhuang 2006). In a CSMA-based MAC protocol, each sender first senses the wireless channel. If the channel is sensed free, the sender transmits its traffic to its receiver; otherwise, the sender defers its transmission. In a busy tone-based MAC protocol, when a sender is transmitting to its receiver, either the sender or its receiver sends a busy tone in another wireless band, to notify other potential senders. Other senders that hear the busy tone will defer their transmissions. It is challenging to guarantee quality of service (QoS) in a CSMA-based MAC protocol or a busy tone-based MAC protocol, due to the following reasons: First, both CSMA- and busy tone-based MAC protocols use channel contention. Second, mobility of the nodes makes it difficult to apply some QoS provisioning techniques such as reservation.

Key Applications

For a wireless ad hoc network, if the locations of the nodes are fixed (such as in a wireless mesh backbone), the work in Jiang et al. (2007) proposes a method to take advantage of the fixed locations of the nodes to improve the QoS, as follows.

The method is based on code division multiple access (CDMA), which allows concurrent transmissions for multiple sender-receiver pairs in a neighborhood. In the network, a common code is available for all nodes, and each node is also assigned with a unique transmitting code and a unique receiving code. As CDMA communica-

tions are interference-limited, one key issue is to make sure that the interference level at each receiver is under control such that the received signal-to-interference-plus-noise ratio (SINR) is not less than a predetermined threshold.

In the system, time is partitioned into slots. At the first portion of each slot, the receiver of each admitted communication link measures the interference level that it receives. At the remaining portion of each slot, if a node, say node i , wants to set up a communication link to its receiver denoted node j , node i first sends a probe message using the common code with a small power level (the purpose of the small power level is not to affect reception of existing links). By reception of the probe message, the receiver of an existing link can estimate the expected extra interference level it will experience if node i transmits its data traffic with normal power level. If the interference level is not tolerable, then the existing receiver sends a busy tone in another frequency band, which will notify node i not to transmit. If node i does not hear a busy tone, which means that all existing links are fine with node i 's transmission, node i sends a request to node j using node j 's receiving code. If node j has an acceptable SINR, it sends back an acknowledgement using its transmitting code to node i , which means that the link from node i to node j is admitted with resources reserved. In other words, node i can keep transmitting to node j by using its transmitting code with transmission quality guaranteed.

The advantages of the method are as follows: First, it is distributed, and thus, a central station is not needed. Only local observation of a node is needed for the node to decide whether or not its link can be admitted. Second, resource reservation can be achieved. Once a link is admitted, it does not need channel contention as in CSMA- or busy tone-based MAC protocols. Rather, the link admission means that QoS of all existing links and the new link can be guaranteed. Due to the fixed locations of the nodes, there is no "sudden" interference from moving nodes. Thus, all nodes can enjoy their target data transmission rates within their link durations. This also means low overhead, as the overhead only happens in the

link admission phase. Third, adaptive transmission rate can be achieved. Note that each receiver is aware of its received SINR. If the received SINR of a receiver is larger than its target SINR value, the receiver can notify its sender to transmit with a higher data rate by using variable spreading gain CDMA or multi-code CDMA. Fourth, each existing receiver estimates possible extra interference level from potential new transmissions. Thus, the interference awareness is receiver-centric. Compared to previous sender-centric schemes (Monks et al. 2001; Muqattash and Krunz 2003) in which each existing link notifies their neighbors of its tolerable interference level, the receiver-centric scheme does not need such notifications and thus, has much less overhead.

Cross-References

- [Interference-Aware Scheduling in Wireless Networks](#)
- [QoS-Aware MAC](#)

References

- Haas ZJ, Deng J (2002) Dual busy tone multiple access (DBTMA) – a multiple access control scheme for ad hoc networks. *IEEE Trans Commun* 50(6):975–985
- Jiang H, Wang P, Zhuang W, Shen X (2007) An interference aware distributed resource management scheme for CDMA-based wireless mesh backbone. *IEEE Trans Wirel Commun* 6(12):4558–4567
- Kleinrock L, Tobagi F (1975) Packet switching in radio channels: part I—carrier sense multiple-access modes and their throughput-delay characteristics. *IEEE Trans Commun* 23(12):1400–1416.
- Monks JP, Bharghavan V, Hwu W-MW (2001) A power controlled multiple access protocol for wireless packet networks. In: *Proceeding of IEEE INFOCOM*, pp 219–228
- Muqattash A, Krunz M (2003) Power controlled dual channel (PCDC) medium access protocol for wireless ad hoc networks. In: *Proceeding of IEEE INFOCOM*, pp 470–480
- Wang P, Zhuang W (2006) An improved busy-tone solution for collision avoidance in wireless ad hoc networks. In: *Proceeding of IEEE ICC*, pp 3802–3807

Interference-Aware Scheduling in Wireless Networks

Yu Wang
University of North Carolina at Charlotte,
Charlotte, NC, USA

Synonyms

[TDMA link scheduling](#); [Wireless link scheduling](#)

Definitions

Interference-aware scheduling determines which links should transmit at which time slots so that all packets transmitted by the scheduled transmitters can be successfully received at receivers under wireless interference constraint. The successful reception of a transmission depends on several parameters, such as the transmission power, the received signal strength, and the interference caused by simultaneous transmissions. Thus, interference-aware scheduling has to consider interference models, traffic patterns, and network topologies. In addition, it is also often studied jointly with routing and power control. Interference-aware scheduling is one of the most fundamental problems for performance optimization in wireless networks.

Background

Wireless network systems have evolved with larger coverage area, higher user density, more diverse applications, and more aggressive usage of spectral resources. While spectral channels are reused for capacity improvement, interference among co-channel transmissions may also cause performance degradation, especially, in multihop wireless networks. One of the solutions is the scheduling of transmissions in a fair and efficient manner to avoid collisions (caused by interference). Such scheduling algorithm is

adopted at link layer in wireless networks as a part of medium access control protocol. Time-division multiple access (TDMA) is one of the most dominant access control solutions to achieve efficient interference-aware scheduling. In TDMA, time is divided into time slots, and a time slot is usually defined as the time unit to transmit a packet between two adjacent nodes. If two nearby transmissions occur simultaneously, collision of packet may happen at the receivers due to interference. TDMA scheduling aims to schedule such transmissions at different time slots to avoid collisions.

Scheduling algorithms in wireless networks have been studied for decades. Depending on the entity that is scheduled, there are node scheduling and link scheduling. In node scheduling, each node is scheduled to transmit at certain time slots, and each node's transmissions have to be received successfully by all its neighbors. In link scheduling, the transmission is intended between two endpoints of a link and have to be received successfully by the receiving node. Each link is scheduled for transmission at certain time slots to avoid interference. In this article, we only focus on wireless link scheduling (more specifically, TDMA link scheduling) in wireless networks. If the reader is interested in node scheduling, please refer to a nice survey by Gore and Karandikar (2011).

Interference-aware link scheduling algorithms have drawn significant attractions from both wireless networking and graph theory communities recently. There are numerous results obtained for various optimization problems under different interference models or network scenarios. Within this article, we briefly discuss the different models and classification methods of wireless link scheduling, summarize a few recent results, and then point out key applications of link scheduling.

Foundations

To perform interference-aware scheduling, first we need to understand the interference model which describe the effects of interference on the

operation of the network. Better characterization of interference can lead to better design of scheduling algorithm and better performance in real environment. There are mainly two interference models: *Protocol interference model* and *physical interference model*.

- *Protocol interference model*. In this model, a transmission by a node v_i is successfully received by a node v_j if and only if the intended destination v_j is sufficiently apart from the source of any other simultaneous transmission, i.e., $\|v_k - v_j\| \geq (1 + \eta)\|v_i - v_j\|$ for any node $v_k \neq v_i$. Here $\|v_i - v_j\|$ is the Euclidean distance between v_i and v_j , and constant $\eta > 0$ models situations where a guard zone is specified by the protocol to prevent a neighboring node from transmitting on the same channel at the same time. This model *implicitly* assumed that each node v_k will adopt the power control mechanism when it transmits signals. A variation of this model is introduced by Wang et al. (2006, 2008) with an assumption that each node v_k has its own fixed transmission power and thus a fixed transmission range t_k . In such model, each node v_k has an *interference range* r_k such that any node v_j will be interfered by the signal from v_k if $\|v_k - v_j\| \leq r_k$ and node v_k is sending signal to some node other than v_j . In other words, the transmission from v_i to v_j is viewed successful if $\|v_k - v_j\| > r_k$ for every node v_k transmitting in the same time slot using the same channel. Here $r_k > t_k$, and we usually assume that $\frac{r_k}{t_k}$ is bounded by a constant larger than one. Protocol interference model is the simplest communication model considering the interference among nodes; it does provide some good estimations of interference and enables a theoretical performance analysis of a number of protocols designed in the literature. However, it is sometimes too simple to capture the complexity of interference (such as the aggregate effect of interference).
- *Physical interference model* (also called *SINR Model*). In this model, node v_j can correctly receive the signal from the sender v_i if and

only if, given a constant $\eta > 0$, the signal-to-interference-plus-noise ratio (SINR) is

$$\frac{P_i \cdot \|v_i - v_j\|^{-\beta}}{N_0 + \sum_{k \in I} P_k \cdot \|v_k - v_j\|^{-\beta}} \geq \eta.$$

Here $\beta > 2$ is the path loss exponent, $N_0 > 0$ is the background Gaussian noise, I is the set of actively transmitting nodes when node v_i is transmitting, and P_i is transmission power of v_i . Note that the physical interference model is based on the power captured at the receiver and depends on the transmission power at all involved nodes. Thus, when we consider interference-aware link scheduling under this model, different power settings can lead to different problems. The common settings include uniform power setting (the transmission power is fixed at the same level for all nodes) or obvious power setting (the transmission power of a node is proportional to the link's path loss factor).

Comparing these two models, physical interference model considers the aggregate effect of interference at the receiver, due to other transmitters, while the protocol interference model focuses on the interference from nearby transmitters. Since the latter is much simple, it has been extensively adopted in the design and analysis of scheduling algorithms. Until recently, the physical interference model becomes more popular. There are also other types and variations of interference models; please refer to a nice survey on interference models (Cardieri 2010).

Besides the interference model, *network model* and *traffic model* are also important aspects to formally define the link scheduling problem. *Network model* includes: (1) What is the network topology (such as tree, mesh, arbitrary, or randomly distributed)? and (2) Does the network node have multiple channels or multiple radios (such as multiple-channel multiple-radio networks, cognitive radio networks)? The *Traffic model* describes the traffic patterns (where the traffics are injected to the network; Are they uniform?) and communication goals (such as

broadcast, point-to-point, multicast, and data collection). Various network/traffic models will change the link scheduling problem and lead to quite different solutions. For example, with multiple channels and multiple radios, the capacity of overall system can be significantly increased, but the search space for optimal scheduling is also enlarged. Scheduling for data collection in a sensor network may restrict the underlying topology to a data collection tree, where a link scheduling could be performed relevantly easily.

The link scheduling problem could have different optimization objectives, e.g., *throughput optimization*, *minimum scheduled length*, *minimum average delay*. For example, *shortest link scheduling* aims to find a link schedule of shortest length for a given subset of links, such that each link (a sender-receiver pair) is scheduled once. *Throughput optimization-based scheduling* (related to capacity problem or called single-slot scheduling) aims to find the maximum number of links that can communicate simultaneously. There are also other variations, e.g., weighted versions, fairness consideration, or under different communication scenarios. Most of these link scheduling optimization problems have been proved to be NP-hard; the majority of research works propose heuristics or approximation algorithms to find their solutions.

The proposed link scheduling algorithms can also be classified into *centralized* and *distributed*. *Centralized algorithms* have the global information of the network and solve the scheduling problem at a controller, while *distributed algorithms* perform the computing distributedly among the nodes, and each node may not have the global information. Usually centralized algorithms can achieve better performances. However, distributed algorithms are preferred at certain wireless networks, such as infrastructure-less multihop wireless networks.

Since there are so many different settings over various models, we would not be able to review all existing solutions for them. Here, we only briefly review a few recent results as examples. For more detailed review, please refer to the survey by Gore and Karandikar (2011).

Various TDMA link scheduling problem under *protocol interference models* have been well-studied, since it has a simple binary interference model. The conflict of transmissions on two distinct links is predetermined independently of the concurrent transmissions at other links. Therefore, the interference relationships can be modeled by a conflict graph, where there is an edge between two nodes (representing two links in the original communication graph) if and only if these two links are interfered to each other under protocol interference model. Then the link scheduling problem is converted to coloring problem on this conflict graph. Various graph theoretical approaches (such as classical greedy algorithm) can then be applied to find good solutions. For example, Wang et al. (2006) studied the TDMA link scheduling under fixed power protocol interference model. They proposed both centralized and distributed algorithms to schedule links over conflict graphs which use time slots within a constant factor of the optimum. Wan et al. (2011a, 2013) investigated wireless link scheduling to maximize concurrent multiframe in multiple-channel multiple-radio wireless networks under protocol interference model. Centralized approximation algorithms are proposed to achieve constant approximation, and both solutions leverage novel conflict graph construction and corresponding scheduling algorithms. Pan et al. (2012) considered link scheduling with cooperative communication in cognitive vehicular networks, where multiple bands and cooperative communication are used. They modeled the problem with a 3D cooperative conflict graph and proposed a heuristic pruning algorithm to solve the overall optimization problem.

The *Physical interference model* captures the aggregate effect of interference in the network; thus the interference at a receiver depends on all concurrent transmissions during that particular time slot, which makes it impossible to generate a conflict graph a priori. Therefore, previous graph coloring techniques based on conflict graphs become inapplicable, and designing scheduling algorithms under physical inter-

ference model becomes especially challenging. Recently, there are significant studies on link scheduling in the physical interference model. For example, Brar et al. (2006) considered the wireless link scheduling under physical interference model and proposed a heuristic algorithm, Greedy Physical, to provide feasible conflict-free scheduling. They also proved that it has an approximation factor if the network is uniformly distributed and every node has the same transmission power. Santi et al. (2009) considered link scheduling under a variation of physical interference model, called graded SINR model, which allows use of “imperfect links” with degraded performance. In addition to conflict-free constraint, the feasible link scheduling also requests the minimum quality of the usable links. Then a constant approximation algorithm (graded SINR) is given under graded SINR model. Blough et al. (2010) also investigated wireless link scheduling under physical interference model with uniform power settings and proposed the first constant approximation algorithm for physical interference model. Wan et al. (2011b) built a unified algorithmic framework based on a maximum independent set of links and developed a set of approximation algorithms for link scheduling with or without power control under physical interference model. Zhou et al. (2014) considered link scheduling to maximize the throughput under physical interference model and proposed the first localized algorithm that can achieve a constant fraction of the optimal capacity. Yu et al. (2016) also studied the distributed link scheduling algorithms with length-monotone sublinear power settings under physical interference model. For dense networks, their solution can achieve a constant approximation.

Besides considering link scheduling problem along, there are also many works combining it with other network issues, such as *routing* (Wang et al. 2008; Badia et al. 2008) and *power control* (Li and Ephremides 2005; Tang et al. 2006), to formulate the optimization problem. Usually, the joint optimization problems become harder to solve; therefore, more complex techniques such as linear programming are often used.

Key Applications

Interference-aware scheduling algorithms are fundamental in every multihop wireless network. They can be applied to improve the overall throughput, reduce packet delay, and increase system capacity. Besides the general multihop wireless networks, interference-aware link scheduling has been adopted in *wireless mesh networks* (Brar et al. 2006; Badia et al. 2008; Gore and Karandikar 2011) (possibly with multiple channels and multiple radios), *wireless sensor networks* (Chen et al. 2011, 2012; Chen and Wang 2012) (for data collection, aggregation, selection), and *vehicular networks* (Pan et al. 2012) (with high mobility, cognitive radio, cooperative communication, etc.).

Cross-References

- ▶ [Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space](#)
- ▶ [Interference-aware distributed resource allocation](#)
- ▶ [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)

References

- Badia L, Erta A, Lenzini L, Zorzi M (2008) A general interference-aware framework for joint routing and link scheduling in wireless mesh networks. *IEEE Netw* 22(1):32–38
- Blough DM, Resta G, Santi P (2010) Approximation algorithms for wireless link scheduling with SINR-based interference. *IEEE/ACM Trans Netwo* 18(6):1701–1712
- Brar G, Blough DM, Santi P (2006) Computationally efficient scheduling with the physical interference model for throughput improvement in wireless mesh networks. In: *Proceedings of the 12th annual international conference on mobile computing and networking (MobiCom'06)*. ACM, New York, pp 2–13
- Cardieri P (2010) Modeling interference in wireless ad hoc networks. *IEEE Commun Surv Tutor* 12(4):551–572
- Chen S, Wang Y (2012) Data collection capacity of random-deployed wireless sensor networks under

- physical models. *Tsinghua Sci Technol* 17(5):487–498
- Chen S, Wang Y, Li XY, Shi X (2011) Capacity of data collection in randomly-deployed wireless sensor networks. *Springer Wirel Netw (WINET)* 17(2):305–318
- Chen S, Huang M, Tang S, Wang Y (2012) Capacity of data collection in arbitrary wireless sensor networks. *IEEE Trans Parallel Distrib Syst (TPDS)* 23(1): 52–60
- Gore AD, Karandikar A (2011) Link scheduling algorithms for wireless mesh networks. *IEEE Commun Surv Tutorials* 13(2):258–273
- Li Y, Ephremides A (2005) Joint scheduling, power control, and routing algorithm for ad-hoc wireless networks. In: *Proceedings of the 38th annual Hawaii international conference on system sciences*
- Pan M, Li P, Fang Y (2012) Cooperative communication aware link scheduling for cognitive vehicular networks. *IEEE J Sel Areas Commun* 30(4):760–768
- Santi P, Maheshwari R, Resta G, Das S, Blough DM (2009) Wireless link scheduling under a graded SINR interference model. In: *Proceedings of the 2nd ACM international workshop on Foundations of wireless ad hoc and sensor networking and computing*. ACM, pp 3–12
- Tang J, Xue G, Chandler C, Zhang W (2006) Link scheduling with power control for throughput enhancement in multihop wireless networks. *IEEE Trans Veh Technol* 55(3):733–742
- Wan PJ, Cheng Y, Wang Z, Yao F (2011a) Multiflows in multi-channel multi-radio multihop wireless networks. In: *2011 Proceedings IEEE INFOCOM*, pp 846–854. <https://doi.org/10.1109/INFOCOM.2011.5935308>
- Wan PJ, Frieder O, Jia X, Yao F, Xu X, Tang S (2011b) Wireless link scheduling under physical interference model. In: *2011 Proceedings IEEE INFOCOM*, pp 838–845
- Wan PJ, Jia X, Dai G, Du H, Wan Z, Frieder O (2013) Scalable algorithms for wireless link schedulings in multi-channel multi-radio wireless networks. In: *Proceedings IEEE INFOCOM*. IEEE, pp 2121–2129
- Wang W, Wang Y, Li XY, Song WZ, Frieder O (2006) Efficient interference-aware TDMA link scheduling for static wireless networks. In: *12th ACM annual international conference on mobile computing and networking (MobiCom 2006)*
- Wang Y, Wang W, Li XY, Song WZ (2008) Interference-aware joint routing and TDMA link scheduling for static wireless networks. *IEEE Trans Parallel Distrib Syst (TPDS)* 19(12):1709–1726
- Yu D, Wang Y, Hua Q, Yu J, Lau FC (2016) Distributed wireless link scheduling in the SINR model. *J Comb Optim* 32(1):278–292
- Zhou Y, Li XY, Liu M, Mao X, Tang S, Li Z (2014) Throughput optimizing localized link scheduling for multihop wireless networks under physical interference model. *IEEE Trans Parallel Distrib Syst* 25(10):2708–2720

Inter-flow Network Coding

- [Joint Network Coding and Opportunistic Forwarding](#)

Internet of Medical Things

Mohammad Upal Mahfuz

Richard J. Resch School of Engineering,
University of Wisconsin-Green Bay, Green Bay,
WI, USA

Synonyms

[Biomedical Internet of Things](#); [Healthcare Internet of Things](#); [Internet of Things \(IoT\)](#); [Medical Internet of Things](#)

Acronyms

IoT Internet of Things

IoMT Internet of Medical Things

Definition

The Internet of Medical Things (IoMT) is defined as a healthcare application of the Internet of Things (IoT) that enables the use of connected medical devices for sensing physiological data, storing medical information and/or records, and providing healthcare services by using sensor and actuator networks and the Internet.

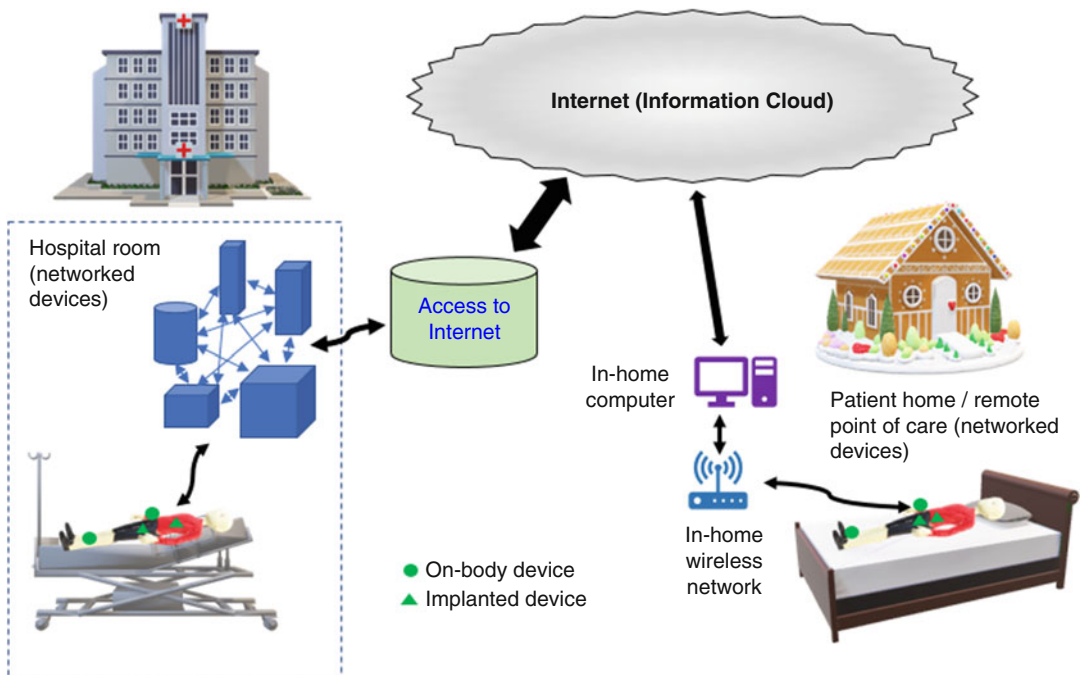
Key Points

The IoT is a novel technology paradigm that has gained a huge interest in the field of wireless telecommunications in the last decade. The IoT is a network of things or physical objects that promises to connect hundreds of billions of things

or physical objects to the Internet (Jha 2018). The fundamental principle of the IoT is that things or physical objects, e.g., radio-frequency identification (RFID) tags; sensors and actuators of various types, sizes, and capabilities; data processing and storing elements; health-monitoring elements of the physical objects; intelligent decision-making and optimizing elements; computers and phones; etc., in the physical environment would use information and communications technologies to connect to each other and share information in order to achieve common goals. The idea of the IoT has been quite investigated in the last decade leading to potential applications for the emerging society (Al-Fuqaha et al. 2015; Atzori et al. 2010; Gubbi et al. 2013; Yin et al. 2018; Zanella et al. 2014; Li et al. 2018; Akmandor and Jha 2018; Yin et al. 2016; Russo and Eriksson 2018; He et al. 2018; Charlotte 2017; Coburn 2016; Marr 2018).

The IoMT is a major application area of the IoT targeted for the healthcare industry toward the smart healthcare paradigm. In the IoMT set-

ting, all the medical devices, ranging from stand-alone macroscale devices and equipment located at the hospitals and healthcare service providers, in-home medical devices and point-of-care units for remote point-of-care applications, and smart wearable devices (sensors, actuators, and communications devices) to mobile healthcare applications and biomedical implants, would potentially be connected to the Internet, which would provide services to personalized healthcare needs and hence a higher standard of patient care in the offered services by accessing medical information and healthcare big data (Meiling 2015). The IoMT paradigm has opened the door to a huge opportunity to receive healthcare services where the healthcare professionals and the patients could be at remote locations and thus realize remote point of care for elderly people and patients with limited mobility. Figure 1 shows a typical IoMT system where patients can be connected to healthcare service providers from patient homes or remote point-of-care as well as hospital rooms.



Internet of Medical Things, Fig. 1 A typical IoMT system

In general, the IoMT would possess similar structural elements as those of the IoT, namely, identification, sensing, communication, computation, services, and semantic elements (Al-Fuqaha et al. 2015). The key enabling technologies of the IoMT would be similar to those of the IoT, namely, identification tags and tag readers, e.g. RFID tags and readers, sensor networks, and a software layer, called the middleware, introduced between the technology and application levels (Atzori et al. 2010). Both wired and wireless communication systems and networks would be applicable to the IoMT framework. In addition to real-time access to IoMT data (information) clouds, all the data would be available to the healthcare service providers on non-real-time basis by accessing the medical data storage systems providing opportunities for post-analysis of physiological signals.

Biomedical implants and wearable medical devices would potentially play an important role in the IoMT because in the last decade their computational power, capacity, and networking capability have improved significantly, while their sizes have decreased opening the doors to smaller devices with enhanced performances (Yin et al. 2018). To this end, communication and networking among biomedical implants, from implants to on-body sensors (wearable devices) and from on-body wearable devices to stand-alone medical devices connected to the Internet, would open new research avenues within the IoMT area.

IoMT offers many benefits toward smart healthcare systems; however, major applications of the IoMT can be grouped into identification, authentication, and tracking of medical objects and people, sensing and monitoring of physiological signals, and collecting and storing of medical data and later sharing with appropriate parties when necessary (Atzori et al. 2010). The IoMT also provides opportunities for emerging medical applications, namely, remote monitoring of patients allowing for a significant cost saving by minimizing the number of hospital visits especially for the elderly patients, remote point of care at the patient locations, and remote surgery.

While the IoMT also offers other promises including better operational efficiency through technologically enhanced automation in health-care monitoring and access to medical big data, it may also bring along some risks associated with its technical implementation and medical data handling. For instance, security and privacy in medical data transfer and the consequences of network connection failures could be considered vital factors in the implementation of the IoMT in the near future.

Cross-References

- Cloud Storage Security
- Data Analytics
- Dedicated Short-Range Communication (DSRC)
- IoT Security
- Machine-Type Communication
- Narrowband Internet of Things
- Privacy-Preserving Healthcare Service
- Smart Device-to-Device Communication: Communication Establishment and Maintenance
- Wireless Data Security
- Wireless Sensor Network Security

References

- Akmandor AO, Jha NK (2018) Smart health care: an edge-side computing perspective. *IEEE Consum Electron Mag* 7(1):29–37
- Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M (2015) Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Commun Surv Tutor* 17(4):2347–2376
- Atzori L, Iera A, Morabito G (2010) The Internet of things: a survey. *Comput Netw* 54(15):2787–2805. Elsevier
- Charlotte H (2017) The Internet of medical things: monitoring and improving performance, website: <https://www.hcinnovationgroup.com/cybersecurity/internet-of-things-iot/article/13008119/the-internet-of-medical-things-monitoring-and-improving-performance>. Accessed 31 July 2019

- Coburn KR (2016) The Internet of medical things. *Scitech Lawyer* 12(3):18–20
- Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comput Syst* 29(7):1645–1660
- He D, Ye R, Chan S, Guizani M, Xu Y (2018) Privacy in the Internet of things for smart healthcare. *IEEE Commun Mag* 56(4):38–44
- Jha NK (2018) Smart healthcare. 2018 IEEE international conference on consumer electronics (ICCE). 12–14 Jan 2018
- Li S, Da Xu L, Zhao S (2018) 5G Internet of things: a survey. *J Ind Inf Integr* 10:1–9. Elsevier
- Marr B (2018) Why the Internet of medical things (IoMT) will start to transform healthcare in 2018, Forbes, website: <https://www.forbes.com/sites/bernardmarr/2018/01/25/why-the-internet-of-medical-things-iomt-will-start-to-transform-healthcare-in-2018/#587ef4194a3c>. Accessed 31 July 2019
- Meiling B (2015) Internet of medical things to change care model, compensation. *San Diego Bus J (Health Care)*, 7 Sept: 4
- Russo N, Eriksson J (2018) The Internet of things and people in healthcare. In: Hassan QF (ed) *Internet of things A to Z: technologies and applications*, 1st edn. Wiley, Hoboken, pp 447–473
- Yin Y, Zeng Y, Chen X, Fan Y (2016) The Internet of things in healthcare: an overview. *J Ind Inf Integr* 1: 3–13. Elsevier
- Yin H, Akmandor AO, Mosenia A, Jha NK (2018) Foundations and trends in electronic design automation. *Smart Healthcare* 12(4):401–466. Now Publishers
- Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities. *IEEE Internet Things* 1(1):22–32. <https://doi.org/10.1109/JIOT.2014.2306328>

Internet of Things

- ▶ [Machine-Type Communication](#)
- ▶ [Ubiquitous Big Data](#)

Internet of Things (IoT)

- ▶ [Anomaly Detection for IoT Systems](#)
- ▶ [Architecture and Data Management for Smart Community Information Platform](#)
- ▶ [Internet of Medical Things](#)

Internet of Things (IoT) Forensic Science

Eyhab Al-Masri

School of Engineering and Technology,
University of Washington Tacoma, Tacoma, WA,
USA

Synonyms

[Edge forensics](#); [Fog forensics](#); [IoT forensic science](#); [IoT forensics](#)

Definition

IoT Forensic Science (IoTFS) is the use of digital forensic processes and procedures relating to the recovery of digital evidence originating from one of more IoT devices toward preserving, identifying, extracting, documenting, interpreting, and presenting digital evidence for the purpose of reconstructing IoT-related events. IoT-related events may reside across one or more configurable computing resources that are in close proximity to where an event has occurred (i.e., fog- or edge-based) or reside on distributed environments (i.e., private, public, or hybrid clouds). IoT forensic science extends that of digital forensics with main focus on the reconstruction of events involving one or more IoT devices to obtain a digital evidence.

IoT Digital Evidence (IoTDE) includes data or events that can be derived from a set of identifiable and configurable IoT objects including IoT devices, nodes, gateways, or cloud resources comprising of a variety of digital media including text files, digital multimedia files (e.g., audio, video, and images), databases, computer programs, hardware resources (e.g., hard disks, other disk drives, SIM cards, keystroke logger media, volatile memory).

Historical Background

Early generations of computers (between 1960s and 1980s) were primarily owned and operated by large corporations, universities, government agencies, and research centers (Sachowski 2019). The security of these systems was mainly the responsibility of system administrators managing the daily operations of these computers (Sachowski 2019). A number of government agencies including the US Department of Defense, US Internal Revenue Service (IRS), and Federal Bureau of Investigation (FBI) began forming ad hoc groups that volunteered to train case investigators on gathering information or extracting evidence from computer systems, mainly mainframe computers (Pollitt 2010). In 1976, Donn Parker, an information security researcher, introduced a book titled *Crime by Computer* describing approaches for using digital information to solve crimes that involved the use of computers.

With the introduction of personal computers (PCs) and as new technologies evolved, law enforcement agencies began introducing laws that address new forms of digital crimes. The Florida Computer Crimes Act was among the first computer-related crime laws introduced in 1978. Subsequently, other laws were introduced in response to the growing number of computer crimes including the US Federal Computer Fraud and Abuse Act (1984), an amendment of the Australian Crimes Act (1989) to include computer-related offenses, the British Computer Abuse Act (1990), among many others. In 1990, the International Association of Computer Investigative Specialists (IACIS) was formed which provided training services to computer forensics professionals around the world. In 1993, the Federal Bureau of Investigation (FBI) hosted the First International Conference on Computer Evidence. In 1995, the International Organization on Computer Evidence (IOCE) was formed.

A number of forensic tools that targeted specific forensic problems were introduced in early 1990s including Gord Hama's RCMP Utilities, and Steve Choy's IACIS Utilities,

among many others. One of the first known commercial computer forensics program that was used by many law enforcement agencies globally was SafeBack, a DOS-based utility program that primarily focused on creating and restoring images of hard disks of Intel-based computer systems (Stephenson and Gilbert 1999). Throughout the 1990s, more advanced forensic tools were introduced such as the EnCase and Forensic ToolKit (FTK), two of the most popular forensic tools. As computing devices and Internet became more ubiquitous, the area of digital forensics witnessed significant growth in size and breadth.

In 2001, the Digital Forensics Research Workshop (DFRWS) was first held and the term digital forensics was introduced (Palmer 2001). The DFRWS also proposed a set of principles that relate to digital evidence including: (a) identification, (b) preservation, (c) collection, (d) examination, (e) analysis, and (f) presentation (Palmer 2001). In 2006, the United States Courts adopted digital information as a new form of evidence implementing a system called electronic discovery for dealing with digital evidence (Civil Litigation Management Manual 2010). In 2014, the National Institute of Standards and Technology (NIST) Cloud Computing Forensic Science Working Group was formed to discuss the forensic challenges in the cloud computing paradigm (NIST 2014).

In recent years, the Internet of Things (IoT) has become a ubiquitous technology due to the proliferation of powerful IoT devices that can perform a variety of tasks. This emerging paradigm, however, has introduced new challenges in dealing with digital evidence for cases where committed crimes involve the use of IoT devices deployed on local (i.e., fog- or edge-based) or cloud (private, public, or hybrid) environments.

Foundations

Digital forensics is a process that encompasses the acquisition and analysis of any forms of digital artifacts on digital devices which can be

used by forensic investigators as digital evidence. NIST defines digital forensics as “the application of science to the identification, collection, examination, and analysis, of data while preserving the integrity of the information and maintaining a strict chain of custody for the data” (Kent et al. 2006).

A digital forensic investigation (DFI) can be defined as a process that determines and correlate information extracted from a digital evidence for the purpose of establishing factual information that can be used for judicial reviews (McKemmish 1999; Jeong 2006). Traditionally, digital forensic investigations follow a reactive approach in which an investigation is typically conducted to solve a digital crime that has been committed.

Digital evidence comprises of data that resides on computing or mobile devices including, but not limited to, data files, text messages, audio, video, or images. Data associated within a running software or hardware can also be part of a digital evidence. A digital evidence is typically obtained when solving crimes that involve the use of digital artifacts as part of a DFI. Digital forensic investigators use DFI tools and techniques for the purpose of collecting digital evidence that can further be used for reporting during the investigation of crimes that involve the use of digital media.

Furthermore, digital forensic investigators gather and analyze all crime-related digital evidence to draw conclusions about a suspect and devise a coherent sketch or plan of how a crime has been committed. To this extent, investigators use a wide variety of forensic tools to help them create case files for recording all case-specific evidence (e.g., blood samples, fluid, fingerprints, hard drives, and computing devices, among others).

Edmond Locard, one of the pioneers in forensic science, introduced what is known as *Locard Exchange Principle* which states that any contact a person makes with another person, place, or object results in the exchange of physical materials (Chisum and Turvey 2000). In the case of IoT, this principle may still apply. For example, if a suspect attempts to control an IoT device, the exchange of physical materials or evidence

in this case are the data files logging activities that can provide tracing information or clues. However, IoT system attacks may involve remote control in which the suspect may not be present in close proximity of IoT devices being controlled. Hence, the exchange of physical materials in such cases can be limited to only software interactions involving digital, network, or storage media where a physical interaction that can be used as evidence is nonexistent.

Key Applications

There are a number of branches of digital forensic science that span across a variety of disciplines. That is, digital evidence can be extracted from a number of different types of digital devices (e.g., laptops, mobile devices, and hard disks, among many others) that may be deployed across multiple different network environments (e.g., private, public, and hybrid). In forensic science, forensic methodologies that are applied for the gathering of information as part of a forensic investigation are not merely restricted to computing devices and involve other disciplines such as accounting, toxicology, odontology, pathology, psychology, archaeology, and dentistry among many others. The following subsections describe the areas in which digital forensic science is applied with primarily focus on the digital forensic investigation types or branches that involve forensic techniques used to gather evidence from electronic or digital artifacts.

- *Computer forensics* relates to the identification, preservation, collection, analysis, and reporting of digital evidence that can be discovered on computing devices such as computers, laptops, tablets, or storage media (Kruse and Heiser 2001). Examples of storage media include, but not limited to, hard disks, other disk drives, USB drives, BIOS information, and an operating system’s registry hives, among others.
- *Mobile forensics* relates to cases involving mobile device technology (Reiber 2016)

(e.g., mobile phones, game consoles, and GPS devices, among others) where evidence that can be recovered for the purpose of information gathering involves digital artifacts on a mobile device. Examples of digital artifacts include, but not limited to, SIM card, images, documents, contact lists, configuration files, and mobile network information, among others.

- *Network forensics* deals with gathering digital evidence through network traffic (Messier 2016). This process involves monitoring, capturing, storing, and analyzing network traffic for collecting digital evidence or intrusion detection. An example of network-related evidence is network traffic filtration which ensures only authorized users can gain access to parts of a desired network.
- *Memory forensics*, also known as *live forensics or acquisition*, deals with recovering evidence from a live setting involving the gathering, capturing, and analyzing of volatile data where evidence is lost when a device or system is powered off (Ligh et al. 2014). Volatile data may reside, for example, on random access memory (RAM) or as data travels through a network channel.
- *Electronic discovery or e-discovery* is a reporting technique that facilitates an investigation in which gathered information is obtained as electronically stored information (ESI) for the purpose of supporting government investigations or a litigation (Phillips et al. 2013). Examples of e-discovery ESI include, but not limited to, identification of documents, preserving data relevant to an investigation, transferring data to legal counsel, and extracting text or metadata from native documents.
- *Database forensics* deals with the recovery of digital evidence involving the use of databases (Adedayo and Olivier 2014) or a database management system (DBMS). Database forensics is required in scenarios such as database failure, deleting persistent data, and detecting suspicious behavior in using the database, among others. Examples involving the use of database forensics include, but not

limited to, fragmented data, stored metadata, or page information reconstruction.

- *Cloud forensics* deals with the recovery of digital evidence involving a cloud computing resource (NIST 2014). The process includes identifying, gathering, preserving, and reporting elements of a digital evidence that may reside across distributed environments. Cloud forensics is an extension of the network forensics. An example that involves the use of cloud forensics is the process of reconstructing a cloud-based event across a cloud-computing environment.
- *IoT forensics* is an all-encompassing forensic type that utilizes many of the other forensic approaches. IoT forensics deals with the recovery of digital evidence from IoT devices. The process includes identifying, gathering, capturing, preserving, and reporting elements of a digital evidence that may reside across shared pool of computing resources and which reside locally (fog- or edge-based) or on distributed environments (cloud). An example that involves the use of IoT forensics is reconstructing IoT-based events that have occurred within IoT systems (e.g., unauthorized access to control an ultrasonic IoT sensor). IoT forensics is an extension of cloud forensics.

IoT Devices and Digital Artifacts

Over the past few years, digital technologies have played a major role in producing various digital forms of artifacts containing information that can be used or shared over the Internet. As we rely on using digital technologies for the dissemination and processing of information, new challenges and trends in digital crimes are emerging.

There are numerous synonyms that are used interchangeably for the distinct role played by digital technologies relating to crimes such as cybercrime, digital crime, e-crime, computer crime, and high-tech crime, among many others. Donn Parker (1998) studied the various types of crimes that employ computers and categorized them into four different roles: (a) object of crime, (b) subject of crime, (c) tool of crime, and (d)

Internet of Things (IoT) Forensic Science, Table 1 Relating Donn Parker Categorization to the Internet of Things (IoT)

Category	Example of a baby-monitoring IoT system
Object	An Internet-connected baby monitoring device is damaged
Subject	Monitoring device camera is infected with malware (e.g., Mirai)
Tool	Use IoT device to carry a DDoS attack causing outages via cloud infrastructure
Symbol	Attacker uses baby monitoring device to threaten or intimidate users of system

symbol of crime. Although these categories were identified in the 1970s, they are still applicable today across many areas including IoT as presented in Table 1.

However, the categories identified by Donn Parker (1998) do not consider the differences between the type of component that can be an object, subject, tool, or symbol. In some cases, for example, the computing device hardware can be considered as a contraband, instrumentality of crime, or evidence of crime, whereas this hardware serves as a medium for storing files that are themselves contraband, instrumentalities of crime, or evidence of crime (Jarrett and Bailie 2015).

With the increasing number of IoT devices, the need for effective digital forensic techniques that can be tailored to IoT systems for solving computer-related crimes becomes inevitable. Traditionally, digital forensic techniques and tools have been dependent file systems to forensically extract digital evidence, recover data, or reconstruct events. Such techniques are no longer be applicable or effective in the IoT paradigm (Al-Masri et al. 2018).

IoT Forensics

An IoT system typically employs a cloud platform for the deployment and management of IoT services or applications running on remote, virtualized environments that are managed by third party cloud providers. This deployment model poses serious implications on how digital evidence can be obtained. In addition, post-event response activities such as gathering and processing digital evidence generally occur after an incident or a digital crime is committed. As

the number of IoT devices proliferate, applying this reactive process of digital forensics becomes increasingly inadequate. It would be desirable to employ a proactive process such that gathering and preserving digital evidence is transpired prior to the occurrence of an accident. This can be achieved by equipping an IoT system to be forensic ready.

At the core of an IoT system is a device layer that provides the basic functional capabilities such as sensing and actuating. Through this layer, IoT devices can collect, process, or exchange data. In order to perform any of these tasks, an IoT device need to be equipped with sufficient capabilities embedded within the device itself including, for example, computational capability, sensing or actuating, power source, and connectivity, among others. However, IoT devices vary in the degree to which they offer these capabilities. In addition, IoT devices may not be virtually accessible which makes the task of collecting information that can be used as digital evidence challenging.

Furthermore, digital forensic techniques for many years have primarily focused on content and relied on the gathering and tracing of information without considering the context in which devices were used in a malicious attack (Al-Masri et al. 2018). Due to the fact that IoT devices are typically constrained in terms of storage and computational capabilities, it is a very common task that data is overwritten occasionally. In addition, IoT systems consist of heterogeneous devices that vary in terms of hardware types as well as identification, sensing, and communication capabilities. This makes the task for recovering data or collecting digital forensic evidence more challenging. Listed below are some of the key characteristics that require

addressing when considering IoT forensics and dealing with digital evidence involving IoT devices.

- *Aggregation*: IoT services are generally deployed across a distributed network. The ability to reconstruct events that may be disseminated through a number of virtual resources to be used for digital evidence is time consuming and complex.
- *Accessibility*: Gaining access to forensic evidence in cloud environments involves IoT service providers who may be reluctant to share data or enable investigators to perform information gathering.
- *Availability*: In the event a breach occurs within an IoT system, some services may be interrupted or network may become unavailable.
- *Deployment*: The process of acquiring IoT digital evidence varies based on the type of environment in which the IoT system is running. For example, processing or collecting evidence varies when an IoT system is deployed on an Infrastructure-as-a-Service (IaaS) or Platform-as-a-Service (PaaS).
- *Documentation*: The heterogeneity in device types is common in IoT systems which makes the process of documenting digital evidence a very complex task and requires extensive domain knowledge across many technologies and protocols.
- *Confidentiality*: IoT systems may have services or applications deployed on the cloud or fog gateways. This increases the risk of data becoming compromised due to the fact that multiple entities or devices may be involved. In addition, multi-tenancy increases the risk of breaching data confidentiality due to data remanence in IoT environments.
- *Integrity*: IoT systems that reside in multi-tenant environment such that data may be shared by many computing devices in a distributed environment can cause digital evidence to be altered without the ability of accurately tracing information.
- *Nonrepudiation*: IoT devices have limited computational capabilities which constrain

their ability to question, repudiate, or deny a traceable message that may have changed by a malicious attack.

- *Preservation*: In cases cloud service providers provide investigators access to IoT services, an attacker may purposefully remove or destroy tracing information.
- *Privacy*: IoT systems depend on the use of services to execute specific tasks through resource URIs (as in the case of REST). In such cases, a URI may expose private metadata thus breaching confidentiality and privacy as well as it may allow malicious entities to infer some information about an IoT device. For an example, assume a REST service URI with the address `http://myHouse/basement/myOffice/light/1` is used to control the status of a light device that is part of a smart home system. Using this URI, a malicious entity may acknowledge that using a HTTP post request to the device translates into turning on the light in the office located in the basement of a tenant's house. The loss of privacy and confidentiality makes the task of reconstructing events or identifying digital evidence a very complex process.
- *Provisioning*: IoT services' provisioning require elasticity which can scale up or down, thus releasing some capabilities which can cause digital evidence to be changed or an event to be lost during evidence collection.
- *Qualifications*: Certain digital forensic techniques may not be relevant to IoT forensics. Therefore, an investigator needs to have sufficient qualifications and expertise when collecting digital evidence involving an IoT system.
- *Trustworthiness*: It is possible that an IoT device or node gets tampered with, stolen, or loses connectivity within an IoT system. The ability to track devices in such situations is nearly impossible.
- *Tracking*: Prior to information gathering by a digital investigator, there is no guarantee with respect to an investigator being the first person on the scene to collect data (e.g., first responder).

- *Size*: The miniaturization of IoT devices limits their capabilities and as a result these devices may not be equipped with fault-tolerant mechanisms in case of malfunctioning.

Future Directions

Digital forensics that involve IoT systems or IoT devices is in the infancy stage. IoT forensics need to transcend the acquisition of data stored locally on hardware resources which can be used for performing detailed analysis or examination of traffic logs. The fact that IoT devices may not be easily reachable or virtually accessible over the Internet requires more adaptive digital forensic techniques that automate some of the basic digital forensic functions. In addition, generated and stored data play a critical role in the acquisition of digital evidence for IoT systems. Therefore, an IoT forensic model needs to maximize the potential use of digital evidence while minimizing costs associated with forensic investigations (Sachowski 2019).

Maintaining and preserving electronically stored information (ESI) that can be used by an IoT system for the purpose of recording users' activities or interactions are of critical importance to IoT forensics. Examples of such records include authentication records, event logs, error logs, transactional records, database records, registry hives, security logs, and configuration files, among many others. Furthermore, applying automatic machine learning and artificial intelligence algorithms that can identify suspicious activities across any ESI along with other background evidence (e.g., authentication records, network device logs) would render automatic forensic processes and facilitate automatic detection of abnormalities.

It is imperative that IoT systems are equipped with the tools necessary to prepare for the need of information gathering for digital evidence. To this extent, it is of great importance that an organization considers formalized plans which identify key goals that affect digital evidence through a set of useful policies and guidelines. These policies and guidelines will help in creating a framework

that can be used to achieve a high-level of digital forensic readiness.

Digital forensic investigations are generally performed as a response to an incident that involves a digital crime. This technique may not be useful for IoT forensics. Information gathering for a digital evidence must be performed in real-time or detected in advance. To this extent, IoT systems must be designed such that they are digital forensics ready. IoT forensics readiness means that an IoT system has the ability to preemptively exploit electronically stored information for the purpose of collecting credible or admissible digital evidence when signs of a digital crime are evolving or as the residual risks associated with an incident increase. This will enable IoT systems increase the potential use of digital evidence whereas organizations will be able to mitigate the impact of risks proactively while reducing investigative costs.

In order to achieve IoT forensic readiness, a number of practices should be discussed. Listed below are some of the practices or procedures that should be considered by organizations for supporting IoT forensics readiness when developing IoT systems.

- Identify business risks clearly and associate a ranking priority to each risk (higher priority indicates high risk).
- Identify all IoT data sources at the local (e.g., fog- or edge-based) and cloud (e.g., public, private, and hybrid) levels and associate an integrity ranking priority to each data source.
- Associate the risks with identified IoT data sources.
- Document the process involved in information gathering for an IoT digital evidence from IoT devices.
- Develop supporting capabilities that enable the automatic detection of incidents and gathering of digital evidence as incidents evolve. This can be achieved by equipping the system to automatically detect anomalies in data streams or event records through the use of machine learning or artificial intelligence techniques.

- Establish a strategy for detecting and collecting credible or admissible digital evidence.
- Identify the process associated with formalizing digital investigations as a result of detecting incidents, suspicious events, or abnormalities in sensory data streams.
- Establish a strategy for deterring incident events which can be achieved using fault-tolerance techniques.
- Establish a framework for presenting credible evidence based on factual information.

In digital forensic investigations, establishing a framework for identifying and maintaining a credible or admissible digital evidence is critical for yielding accurate conclusions based on factual information. IoT forensics is an all-encompassing apparatus that uses forensic approaches spanning across a number of areas including cloud, computers, database, network, and mobile, among many others. IoT forensics deals with reconstructing events and the recovery of digital evidence involving the use of IoT devices. The proliferation and heterogeneity of IoT devices increase the need for organizations to employ processes, procedures, training, and tools that are specifically designed for digital forensics involving IoT devices while ensuring reliability when drawing conclusions as part of a digital investigation. The growing number of IoT devices also increases the need for organizations to collaborate at a level that supports the interoperability of the processes or procedures applied in IoT forensic investigations.

Cross-References

- [Internet of Things](#)
- [IoT Forensics](#)

References

Adedayo O, Olivier M (2014) Schema reconstruction in database forensics. In: *Advances in digital forensics X*. DigitalForensics, IFIP advances in information and communication technology, vol 433. Springer, Berlin/Heidelberg

- Al-Masri E, Bai Y, Li J (2018) A fog-based digital forensics investigation framework for IoT systems. In: *IEEE international conference on Smart Cloud*, pp 196–201
- Chisum W, Turvey B (2000) Evidence dynamic: Locard's exchange principle & crime reconstruction. *J Behav Profile* 1(1)
- Civil Litigation Management Manual (2010) The judicial conference of the United States Committee on court administration and case management. <https://www.fjc.gov/sites/default/files/2012/CivLit2D.pdf> (Accessed 5 Apr 2020)
- Draft NISTIR 8006: Cloud computing forensic science challenges (2014) NIST cloud computing forensic science working group information technology laboratory. https://csrc.nist.gov/csrc/media/publications/nistir/8006/draft/documents/draft_nistir_8006.pdf (Accessed 5 Apr 2020)
- Jeong R (2006) FORZA – digital forensics investigation framework that incorporate legal issues. *Digit Investig* 3:29–36
- Jarrett H, Bailie M (2015) Prosecuting computer crimes, computer crime and intellectual property section. Criminal Division U.S. Department of Justice. <https://www.justice.gov/sites/default/files/criminal-ccips/legacy/2015/01/14/ccmanual.pdf> (Accessed 5 Apr 2020)
- Kent K, Chevalier S, Grace T, Dang H (2006) Guide to integrating forensic techniques into incident response. <https://www.nist.gov/publications/guide-integrating-forensic-techniques-incident-response> (Accessed 5 Apr 2020)
- Kruse W II, Heiser J (2001) *Computer forensics: incident response essentials*. Pearson Education, London
- Ligh M, Case A, Levy J, Walters A (2014) *The art of memory forensics: detecting malware and threats in Windows, Linux, and Mac memory*. Wiley, New York
- McKemmish R (1999) What is forensic computing? *Trends & issues in crime and criminal justice*, no. 118. Australian Institute of Criminology, Canberra. <https://www.aic.gov.au/publications/tandi/tandi118>
- Messier R (2016) *Network forensics*. Wiley, New York
- Palmer G (2001) A road map for digital forensic research. In: *First digital forensic research workshop*, pp 27–30
- Parker D (1998) *Fighting computer crime: a new framework for protecting information*. Wiley, New York
- Phillips A, Steuart C, Godfrey R, Brown C (2013) *E-discovery: an introduction to digital evidence*. Cengage Learning, Boston
- Pollitt M (2010) A history of digital forensics. In: *International conference on digital forensics*, HAL, pp 3–15
- Reiber L (2016) *Mobile forensic investigations: a guide to evidence collection, analysis, and presentation*. McGraw-Hill Education, New York
- Sachowski J (2019) *Implementing digital forensic readiness from reactive to proactive process*. CRC Press, Boca Raton
- Stephenson P, Gilbert K (1999) *Investigating computer-related crime*. CRC Press, Boca Raton

Internet of Things for Vessels

► Maritime Internet of Vessels

Internet of Things: Architecture, Key Applications, and Security Impacts

Hansong Xu, Fan Liang, and Wei Yu
Department of Computer Engineering, Towson
University, Towson, MD, USA

Introduction

Thanks to recent advances in Internet of Things (IoT) technologies (Lin et al. 2017; Stankovic 2014; Yu et al. 2018), a number of IoT-based systems, including the smart grid, smart transportation, and smart manufacturing, among others, can be supported (Catarinucci et al. 2015; Lin et al. 2015; Wu et al. 2017; Xu et al. 2017a,b, 2018a,b). With IoT, timely monitoring and controlling of physical components via deployment of both sensing and actuating devices can be enabled. Such a feedback loop for monitoring and control can improve the efficiency of power management, travel planning, and manufacturing production, to name just a few. For example, in the smart grid, physical components (power plants, transmission lines, substations, appliances, etc.) can be monitored and controlled, leading to two-way communication of both power and information flows. Smart meters, as typical sensing devices, can detect the demands from consumers and the state of both traditional energy generators and renewable energy resources and transmit sensed information to the operation center. Based on collected sensing information, the operation center can coordinate the energy supply in supporting the demand variation. In this way, efficient and resilient energy management can be realized in the smart grid.

In the smart transportation system, IoT technologies enable the interconnection of vehicles, roadside units (RSUs), and drivers to improve travel efficiency and safety (Lin et al. 2015; Xu et al. 2017b). For instance, traffic information, which is collected and shared among vehicles, can be used to derive optimal routes for drivers. In addition, emergency safety events (collisions, breakdowns, etc.) during travel can be broadcasted to nearby vehicles to alert drivers or to ask for assistance. Similarly, smart manufacturing is envisioned as the next revolution in the manufacturing industry (Davis et al. 2012; Kang et al. 2016; Kusiak 2017; Xu et al. 2018a). In the smart manufacturing system, IoT integrates system automation, intelligent control, and distributed sensing in all appropriate stages of the manufacturing processes within a factory or plant to optimize resource efficiency and productivity (Lu et al. 2016). As an example, in monitoring of the supply chain along with a local manufacturing process, the optimal production efficiency can be achieved through seamless information collection and dissemination among activities, as well as across the supply chain, including raw material management, logistics, production and consumption, and others.

With the massive deployment of smart sensors and actuators, excessive and valuable data is collected, processed, and shared, making IoT-based systems attractive targets for curious and malicious adversaries. Moreover, IoT-based systems suffer vulnerabilities from the perspectives of software, hardware, communication protocols, and so on, which can be leveraged by adversaries to launch attacks. Cyber-attacks on IoT-based systems can be categorized by their impacts on the three-layer architecture (i.e., perception layer, network layer, and application layer). By exploring cyber-attacks on IoT-based systems, the impacts on smart grid, smart transportation, and smart manufacturing can be further investigated. To design effective defensive schemes, the vulnerabilities shall be investigated first, and then the corresponding defensive schemes shall be examined by considering the nature of IoT-based systems and the impact of the attack. Also, the defensive schemes

must be evaluated comprehensively before wide deployment.

In this entry, we make the following contributions:

- We present the smart grid, smart transportation, and smart manufacturing as three key IoT-based systems from the perspective of a three-layer IoT architecture (i.e., perception layer, network layer, and application layer). In the IoT architecture, the perception layer consists of diverse things, the network layer includes a variety of networks and related technologies, and the application layer provides diverse services for IoT-based systems. We also present the information flows that show the data collection and control command dissemination for the three representative IoT-based systems.
- We design a taxonomy to investigate cyber-attacks on IoT-based systems and their impacts from the perspective of the three-layer system architecture. Particularly, in the perception layer, cyber-attacks (e.g., node capture attacks, malicious code injection attacks, and false data injection attacks, among others) are investigated, as well as their impacts, including data tampering, device tampering, and device authentication disruption. Cyber-attacks (denial-of-service (DoS) attacks, spoofing attacks, sinkhole attacks, and wormhole attacks, among others) are considered in the network layer, as well as their impacts (service disruptions, disrupting device authentication, stealing sensitive data, etc.). In the application layer, phishing attacks, malicious worms/virus, and malicious scripts are considered, as well as their impacts such as stealing sensitive data and damage to the system.
- We conduct a case study on the investigation of data integrity attacks and the design of defensive schemes for the smart grid. We also discuss some open issues, including the investigation of cyber-attack risks on IoT-based systems, the design of defensive schemes, and the development of integrated evaluation platforms.

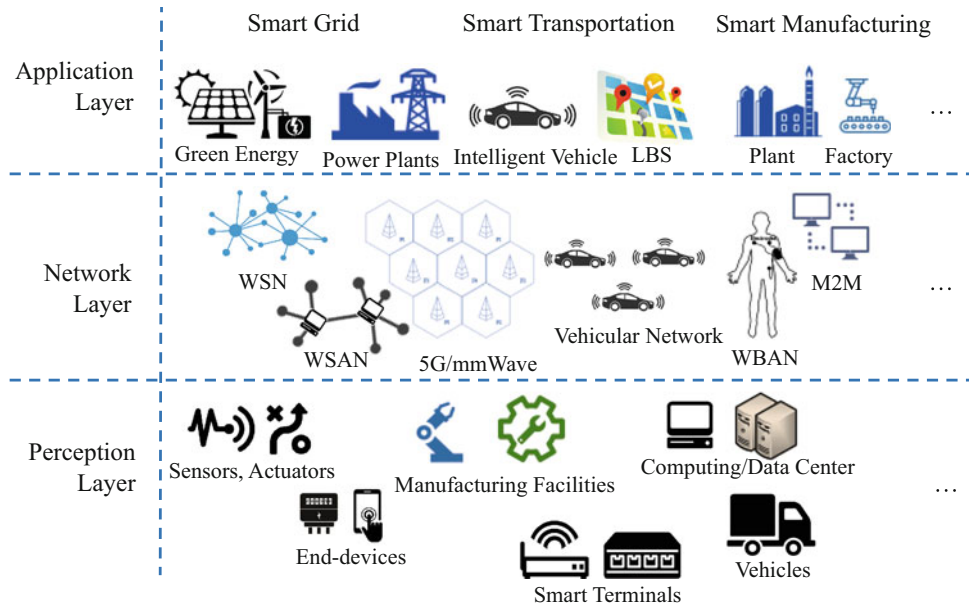
The remainder of the entry is organized as follows: In section “[Overview of IoT](#)”, we present an overview of IoT. In section “[A Taxonomy of Cyber-Attacks on IoT-Based Systems](#)”, we present cyber-attacks on some representative IoT-based systems from the perspective of the three-layer architecture (i.e., perception layer, network layer, and application layer). In section “[Case Study](#)”, we present a case study of cyber-attacks on IoT-based systems, such as data integrity attacks on the smart grid. In section “[Future Directions](#)”, we discuss some open issues from risk assessment, defensive schemes, and evaluation platforms. Finally, we conclude the entry in section “[Final Remarks](#)”.

Overview of IoT

In this section, we first introduce a three-layer IoT architecture and several representative IoT-based systems. We then show how the information flows within the system and introduce the corresponding deployment scenarios for the three representative IoT-based systems.

IoT Architecture and Applications

We can generalize IoT architecture into three primary layers, including the perception layer, network layer, and application layer, as shown in Fig. 1 (Lin et al. 2017). The perception layer integrates things (computing and network-connected devices) in IoT-based systems by direct perception and collection of data associated with the physical objects. The perception layer plays an important role in connecting those things to IoT networks. The network layer consists of a number of IoT networks and related technologies, including wireless sensor network (WSN), wireless sensor and actuator network (WSAN), next generation wireless networks, and so on. After being integrated into the network layer, those IoT networks deliver high volumes of widespread perception data to corresponding recipients. Supported by the network and perception layers, the application layer provides services (e.g., storage, analytics) with the aid of cloud and edge computing infrastructures and machine learning



Internet of Things: Architecture, Key Applications, and Security Impacts, Fig. 1 A three-layer architecture for I-IoT

technologies (Yu et al. 2018; Hatcher and Yu 2018). The smart grid, smart transportation, and smart manufacturing are typical examples of IoT-based systems.

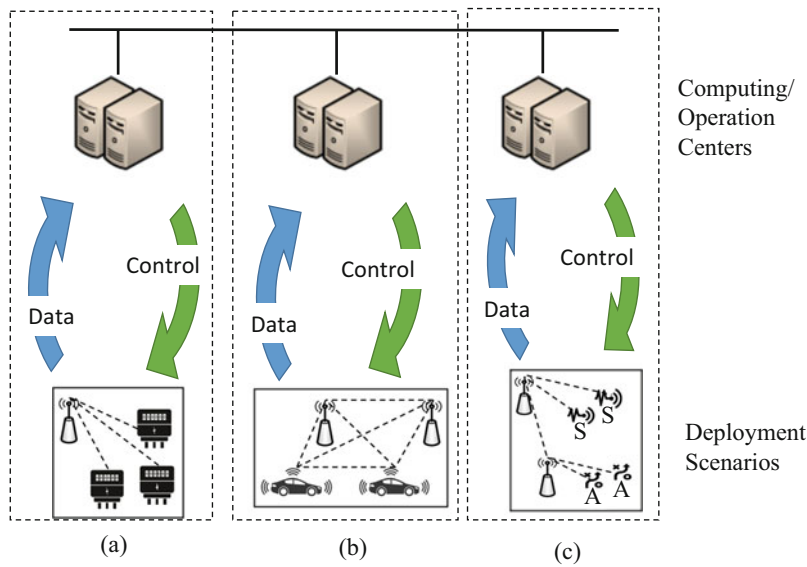
As shown in Fig. 1, the smart grid, smart transportation, and smart manufacturing are based on a vertical architecture. Particularly, the smart grid system integrates cyber-components (i.e., sensing, computation, and communication) in power generation, transmission, and utilization of the physical power grid system, to improve its intelligence and efficiency (Xu et al. 2017a). Similarly, the smart transportation system enables traffic information collection, analysis, and sharing among massively deployed vehicles, RSUs, and travelers, to improve travel management and driving safety (Lin et al. 2015; Xu et al. 2017b). The smart manufacturing system interconnects diverse and multitudinous manufacturing things, including sensors, actuators, controllers, robotic arms, and production lines, among others, to improve productivity and resource efficiency (Xu et al. 2018a). Those IoT-based systems can be generalized as the seamless integration of cyber-components with traditional physical

components from a cyber-physical perspective. In addition to the aforementioned systems, there are many other IoT-based systems, including smart healthcare, smart cities, and smart homes, to name a few.

Feedback Loop Information Flow

We now present the information flow and corresponding deployment scenarios for the aforementioned IoT-based systems in Fig. 2. Generally speaking, in the perception layer, things collect data and push the data upward to the computing operation centers. Then, the computing center leverages powerful computing and analytic tools to process and analyze data and derive optimal decisions. Finally, the center disseminates corresponding control commands and messages to the things (as actuators) in closed-loop.

This closed-loop information flow is generic and can be applied to numerous IoT-based systems. For example, Fig. 2a illustrates an automatic metering scenario in the smart grid. The smart meters collect energy usage data from residential buildings and transmit the data to the corresponding computing operation



Internet of Things: Architecture, Key Applications, and Security Impacts, Fig. 2 Information flow and deployment scenarios for (a) smart grid, (b) smart transportation, and (c) smart manufacturing

center. Figure 2b presents a scenario in which vehicles and RSUs are interconnected for information sharing (e.g., emergency message broadcasting). The traffic information can also be collected and pushed to the computing operation center for optimized traffic management. Figure 2c illustrates a scenario related to sensing and actuation for smart manufacturing. The information can flow among the sensors, controllers, and actuators locally or via the computing operation center when computation-intensive tasks are carried out.

A Taxonomy of Cyber-Attacks on IoT-Based Systems

Based on the layered structure of IoT-based systems and the information flow illustrated in section “Overview of IoT”, massively deployed IoT devices continuously generate and transmit valuable information to support the efficient and resilient operations of physical systems. Nonetheless, this makes IoT-based systems attractive attack targets for adversaries, raising significant security challenges. From

the perspective of the layered architecture, the different security challenges between layers have garnered significant attention from the research community. For example, Lin et al. (2017) investigated cyber-attacks on IoT-based systems from the perspective of the three layers. Node capture attacks, malicious code injection attacks, false data injection attacks, replay attacks, cryptanalysis and side-channel attacks, eavesdropping and interference attacks, and sleep deprivation attacks are considered as representative cyber-attacks in the perception layer. Denial-of-service (DoS) attacks, spoofing attacks, sinkhole attacks, wormhole attacks, man-in-the-middle (MITM) attacks, routing information, Sybil attacks, and unauthorized access are likewise grouped into representative cyber-attacks in the network layer. In addition, the examples of cyber-attacks in the application layer include phishing attacks, malicious worms/virus, and malicious scripts, among others. To understand those cyber-attacks, we investigate the impacts of cyber-attacks on different IoT-based systems in the perception layer, network layer, and application layer in detail.

Cyber-Attacks in Perception Layer

We use Table 1 to illustrate the aforementioned cyber-attacks in the perception layer on the smart grid, smart transportation, and smart manufacturing, as well as their impacts. To be specific, we describe the consequences of node capture attacks (A_1), malicious code injection attacks (A_2), false data injection attacks (A_3), replay attacks (A_4), cryptanalysis and side-channel attacks (A_5), eavesdropping and interference attacks (A_6), and sleep deprivation attacks (A_7) on the smart grid, smart transportation, and smart manufacturing. In addition, we classify those cyber-attacks by their impacts, including data tampering, device tampering, disrupting device authentication, stealing sensitive data, and causing damage.

Regarding node capture attacks (A_1), the adversary could capture, tamper with, and redeploy devices to further gain control over other IoT devices, monitor the system, or introduce incorrect information to the system.

For example, the adversary can first modify smart meters and redeploy them in the smart grid, and then use the compromised smart meters to report false metering data to steal electricity. In addition, the manipulation of a large number compromised smart meters can cause severe economic losses to the system. Similarly, in the smart transportation system, the compromised vehicles could report incorrect traffic status information, which could cause inefficient route management (Lin et al. 2018). An aggressive adversary could use a large volume of compromised vehicles and RSUs to maximize the damage to the transportation infrastructure.

As another example, to launch a false data injection attack (A_3), adversaries inject false data in place of legitimate measurement data to mislead management mechanisms and disrupt power generation, distribution, and utilization. Particularly, the false data injected in the smart grid could maliciously affect a number of critical functions, such as state estimation (Yang et al.

Internet of Things: Architecture, Key Applications, and Security Impacts, Table 1 Cyber-attacks in perception layer and impacts

Attacks in perception layer	Smart grid (Recall Fig. 2a)	Smart transportation (Recall Fig. 2b)	Smart manufacturing (Recall Fig. 2c)
Node capture attacks (A_1) <i>Data tampering</i>	Compromised smart meters report false meter readings	Compromised vehicles or RSUs report false measurement data	Compromised sensors or controllers report false measurement data or false control commands
Malicious code injection attacks (A_2) <i>Device tampering</i>	Adversaries inject malicious code to smart meters to perform illegitimate functions	Adversaries inject malicious code to vehicles or RSUs to conduct illegitimate activities	Adversaries inject malicious code to sensors or controllers to modify control logics
False data injection attacks (A_3) <i>Data tampering</i>	Adversaries inject false data (e.g., meter reading) to steal electricity	Adversaries inject false data (e.g., traffic status) to disrupt traffic management	Adversaries inject false data (e.g., control parameters) to alter control objectives
Replay attacks (A_4) <i>Disrupt device authentication</i>	Grant trust to malicious meters	Grant trust to malicious vehicles or RSUs	Grant trust to malicious sensors, actuators, or controllers
Cryptanalysis attacks and side-channel attacks (A_5) <i>Disrupt device authentication</i>	Adversaries can leverage side-channel information to disrupt the authentication functions	Adversaries disrupt the authentication of vehicles to vehicle and vehicle to RSU	Adversaries disrupt the authentication of manufacturing devices (e.g., robotic arms, machines)
Eavesdropping and interference attacks (A_6) <i>Stole sensitive data</i>	Legitimate data can be stolen by adversaries via eavesdropping	Adversaries eavesdrop sensitive data (e.g., route, destination, identification)	Adversaries eavesdrop sensitive data (e.g., control parameters, productivity)
Sleep deprivation attacks (A_7) <i>Cause damages</i>	Adversaries disrupt the sleep cycle of smart meters	Adversaries disrupt the sleep cycle of RSUs	Adversaries disrupt the sleep cycle of sensors, actuators, and controllers

2014), energy pricing (Lin et al. 2016), and renewable energy resource integration (Lin et al. 2012), among others. In the case of the smart manufacturing system, false data injection attacks can cause severe impacts on manufacturing plants and factories. The input of false data in the closed-loop control system could generate misleading control signals, which negatively impacts production quality, cause damage to manufacturing equipment, and even endanger the safety of workers and the work environment.

Cyber-Attacks in Network Layer

Table 2 illustrates some examples of network-layer cyber-attacks on the smart grid, smart transportation, and smart manufacturing, including DoS attacks (A_8), spoofing attacks (A_9), sinkhole attacks (A_{10}), wormhole attacks

(A_{11}), man-in-the-middle (MITM) attacks (A_{12}), routing information attacks (A_{13}), Sybil attacks (A_{14}), and unauthorized access (A_{15}), as well as their impacts.

For example, the adversary launches DoS attacks to consume all resources for manufacturing control and management. Once successful, the DoS attacks result in the unavailability of critical control services, including process control, product inspection, and so on, greatly hindering or disabling services in the smart manufacturing system. In addition, DoS attacks on the smart grid and smart transportation system could result in the dysfunction of critical services, including power generation, transportation utilization, and pricing, as well as traffic management, location-based services, and driving safety. Moreover, in a large-scale

Internet of Things: Architecture, Key Applications, and Security Impacts, Table 2 Cyber-attacks in network layer and impacts

Attacks in network layer	Smart grid (Recall Fig. 2a)	Smart transportation (Recall Fig. 2b)	Smart manufacturing (Recall Fig. 2c)
DoS attacks (A_8) <i>Service Disruptions</i>	Make IoT system resources unavailable to legitimate requests	Adversaries bombard IoT network, leading to failure on delivery of time-critical tasks from vehicles or RSUs	Adversaries disrupt the timely delivery of time-sensitive control data in control loops
Spoofing attacks (A_9) <i>Disrupt device authentication</i>	The malicious meters could be admitted into the smart grid system	Adversaries use malicious devices to affect the smart transportation system	Malicious devices affect the normal operation of control loops in the smart manufacturing system
Sinkhole attacks (A_{11}) <i>Steal sensitive data</i>	Adversaries can collect a large amount of legitimate smart meter data	Massive sensitive and private vehicle and traffic information can be disclosed	Adversaries can steal sensitive and valuable manufacturing data
Wormhole attacks (A_{10}) <i>Steal sensitive data</i>	Same as A_9	Same as A_9	Same as A_9
MITM attacks (A_{12}) <i>Steal sensitive data</i>	Adversaries can collect and alter legitimate smart meter data via malicious devices	Adversaries can obtain and alter sensitive traffic and vehicle data	Adversaries can obtain and alter sensitive and valuable manufacturing data
Routing information attacks (A_{13}) <i>Cause damages</i>	Disrupting the timely delivery of meter reading data via route loops	Time-critical traffic data is not delivered on time	Time-sensitive control messages fail to be delivered
Sybil attacks (A_{14}) <i>Steal sensitive data</i>	Malicious devices claim several legitimate identities to cause jamming or DoS	Devices cause jamming or DoS in the smart transportation system	Devices cause jamming or DoS in the smart manufacturing system
Unauthorized access (A_{15}) <i>Disrupt device authentication</i>	Adversaries gain unauthorized access to smart meters	Adversaries gain unauthorized access to vehicles and RSUs	Adversaries gain unauthorized access to industrial sensors, actuators, and controllers

Internet of Things: Architecture, Key Applications, and Security Impacts, Table 3 Cyber-attacks in application layer and impacts

Attacks in application layer	Smart grid (Recall Fig. 2a)	Smart transportation (Recall Fig. 2b)	Smart manufacturing (Recall Fig. 2c)
Phishing attack (A_{16}) <i>Steal sensitive data</i>	Adversaries can obtain confidential data from smart meters	Adversaries can obtain sensitive data of vehicles, routes, and traffic in the smart transportation system	Confidential information in the smart manufacturing system is disclosed
Malicious virus/worm (A_{17}) <i>Cause damages</i>	Smart meters are infected by distributed self-propagating malicious software	Infected devices perform illegitimate function to harm the smart transportation system	Propagation and triggering endangers the smart manufacturing services, including smart plants, factories, etc.
Malicious scripts (A_{18}) <i>Cause damages</i>	Adversaries can run malicious scripts in smart meters to execute illegitimate functions	Adversaries run malicious scripts to disrupt smart transportation systems	Adversaries run malicious scripts to disrupt smart manufacturing applications

DoS attacks (e.g., distributed DoS), there exist massive numbers of attacking nodes. Note that how to quickly detect and effectively mitigate DDoS attacks remains an unresolved issue.

Cyber-Attacks in Application Layer

We use Table 3 to illustrate cyber-attacks on the smart grid, smart transportation, and smart manufacturing in the application layer, including phishing attacks (A_{16}), malicious viruses/worms (A_{17}), and malicious scripts (A_{18}), as well as the impacts of the cyber-attacks on the system.

Case Study

To better understand the cyber-attacks in IoT, we now present a case study on data integrity attacks on the smart grid system. Note that our designed methodology can be applied to other IoT-based systems, such as smart transportation system (Lin et al. 2018) and smart manufacturing system (Xu et al. 2018a; Wu et al. 2018).

For numerous reasons, including open network interfaces and the lack of tamper-resistant hardware in SCADA systems in the smart grid, a number of cyber-attacks could be leveraged by adversaries to disrupt the operation of power grid systems. For instance, an adversary could leverage malicious software to launch data integrity attacks, which inject false

data into the smart grid system and disrupt its operations, including the state estimation (Yang et al. 2014), energy pricing (Lin et al. 2016), and renewable and traditional energy resource integrity (Lin et al. 2012), among others. Notice that Stuxnet is one such malicious software that can be used to attack SCADA systems (Langner 2011).

To understand the impact of data integrity attacks on smart grid state estimation and design defensive schemes in response, Yang et al. (2014) addressed two issues: (i) how would an adversary choose the smallest number of smart meters to compromise in order to sufficiently manipulate a given number of states and (ii) how to detect malicious smart meters readings. To tackle the first issue, they formalized the least-effort attack model as an NP-hard problem and designed several heuristic algorithms to derive approximate optimal solutions. Regarding the second issue, they designed two detection schemes, one based on the temporal behavior of compromised meters and the other based on the spatial behavior of compromised meters. To defend against data integrity attacks on the state estimation in the smart grid, Yang et al. (2017) presented a protection-based defensive scheme that deploys highly secured phasor measurement units (PMUs). In this defense, a deployment algorithm was designed to not only derive the smallest number of PMUs that are sufficient to

defeat the attack but also to determine optimal locations for PMU deployment.

Future Directions

As future directions, we shall investigate risks of cyber-attacks, design effective defensive schemes, and leverage integrated evaluation platforms toward security-resilient smart grid, smart transportation, and smart manufacturing systems.

In particular, to investigate risks of cyber-attacks, we shall understand the system vulnerabilities from the perspectives of software, communication, and side-channel, among others (Lin et al. 2012; Yu 2012; Yu et al. 2015). Then, we shall investigate the impacts of system vulnerabilities in terms of key services, operation, and management in disparate IoT systems. Besides, we shall understand the optimal attack strategies (duration of an attack, spatial distribution, and scale, among others) for adversaries, which achieve their attack goals with minimum cost.

We shall study the design of defensive schemes from four stages, such as manufacturing stage, configuration stage, detection stage, and response stage (He et al. 2015; Lin et al. 2013; Yang et al. 2017, 2016). In manufacturing stage, the IoT devices, such as smart meters, RSUs, and controllers, and so on shall be designed and manufactured in a secure way (i.e., minimize vulnerabilities). In the configuration stage, the system configuration, including device deployment, software upgrade, and maintenance, shall be securely conducted. In the detection and response stages, the effective anomaly detection schemes and corresponding responsive plans shall be designed to detect cyber-attacks and mitigate caused damages.

Moreover, we shall leverage integrated evaluation platform to evaluate effectiveness of the defensive schemes. For smart grid, one co-simulation platform is the Fenix framework for Network Co-Simulation (FNCS), which

integrates GridLAB-D and NS-3. The FNCS can capture the interaction between power grid and communication network, as well as computational algorithms and cyber-attacks (Moulema et al. 2015). For the smart transportation, we can leverage integrated testbed, including OMNET++, SUMO, and Veins to describe the interactions between transportation system (SUMO) and communication network (OMNET++) via a Veins framework (Noori 2012). For the smart manufacture system, we can use the wireless cyber-physical simulator (WCPS) to describe the interactions between industrial control systems (Simulink) and communication network (TOSSIM) (Li et al. 2013; Ma et al. 2018; Ma and Lu 2018).

Final Remarks

IoT-based systems, such as the smart grid, smart transportation, and smart manufacturing engender tremendous benefits, including resource efficiency, convenience, productivity, and safety via massive collection, transmission, analysis of valuable sensor, and control information. Despite these benefits, security issues in IoT-based systems caused by vulnerability to cyber-attacks prevent their full realization and thus demand systematic investigation and consideration. In this entry, we first presented an IoT architecture and representative IoT-based systems and described insightful information flows. We then studied cyber-attacks on IoT-based systems and their impacts from the perspectives of the three-layer IoT architecture. We also presented a case study to demonstrate examples of cyber-attacks on the smart grid and discuss several future directions toward securing IoT-based systems.

Acknowledgments This work was supported in part by the US National Science Foundation (NSF) under grants: CNS 1350145. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

- Catarinucci L, De Donno D, Mainetti L, Palano L, Patrono L, Stefanizzi ML, Tarricone L (2015) An IoT-aware architecture for smart healthcare systems. *IEEE Internet Things J* 2(6):515–526
- Davis J, Edgar T, Porter J, Bernaden J, Sarli M (2012) Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Comput Chem Eng* 47:145–156
- Hatcher WG, Yu W (2018) A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access* 6:24411–24432. <https://doi.org/10.1109/ACCESS.2018.2830661>
- He X, Yang X, Lin J, Ge L, Yu W, Yang Q (2015) Defending against energy dispatching data integrity attacks in smart grid. In: 2015 IEEE 34th international performance on computing and communications conference (IPCCC). IEEE, pp 1–8
- Kang HS, Lee JY, Choi S, Kim H, Park JH, Son JY, Kim BH, Do Noh S (2016) Smart manufacturing: past research, present findings, and future directions. *Int J Precis Eng Manuf Green Technol* 3(1): 111–128
- Kusiak A (2017) Smart manufacturing must embrace big data. *Nat News* 544(7648):23
- Langner R (2011) Stuxnet: dissecting a cyberwarfare weapon. *IEEE Secur Priv* 9(3):49–51. <https://doi.org/10.1109/MSP.2011.67>
- Li B, Sun Z, Mechitov K, Hackmann G, Lu C, Dyke SJ, Agha G, Spencer BF (2013) Realistic case studies of wireless structural control. In: 2013 ACM/IEEE international conference on cyber-physical systems (ICCPs). IEEE, pp 179–188
- Lin J, Yu W, Yang X, Xu G, Zhao W (2012) On false data injection attacks against distributed energy routing in smart grid. In: Proceedings of the 2012 IEEE/ACM third international conference on cyber-physical systems, ICCPS'12. IEEE Computer Society, Washington, DC, pp 183–192. <https://doi.org/10.1109/ICCPs.2012.26>
- Lin J, Yu W, Yang X (2013) On false data injection attack against multistep electricity price in electricity market in smart grid. In: 2013 IEEE on global communications conference (GLOBECOM). IEEE, pp 760–765
- Lin J, Yu W, Yang X, Yang Q, Fu X, Zhao W (2015) A novel dynamic en-route decision real-time route guidance scheme in intelligent transportation systems. In: 2015 IEEE 35th international conference on distributed computing systems, pp 61–72. <https://doi.org/10.1109/ICDCS.2015.15>
- Lin J, Yu W, Yang X (2016) Towards multistep electricity prices in smart grid electricity markets. *IEEE Trans Parallel Distrib Syst* 27(1):286–302. <https://doi.org/10.1109/TPDS.2015.2388479>
- Lin J, Yu W, Zhang N, Yang X, Zhang H, Zhao W (2017) A survey on internet of things: architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J* 4(5):1125–1142
- Lin J, Yu W, Zhang N, Yang X, Ge L (2018) Data integrity attacks against dynamic route guidance in transportation-based cyber-physical systems: modeling, analysis, and defense. *IEEE Trans Veh Technol* 67(9):8738–8753. <https://doi.org/10.1109/TVT.2018.2845744>
- Lu Y, Morris KC, Frechette S (2016) Current standards landscape for smart manufacturing systems. *J Res Natl Inst Stand Technol* 8107:39
- Ma Y, Lu C (2018) Efficient holistic control over industrial wireless sensor-actuator networks. In: 2018 IEEE international conference on industrial internet (ICII). IEEE, pp 89–98
- Ma Y, Gunatilaka D, Li B, Gonzalez H, Lu C (2018) Holistic cyber-physical management for dependable wireless control systems. *ACM Trans Cyber-Phys Syst* 3(1):3
- Moulema P, Yu W, Griffith D, Golmie N (2015) On effectiveness of smart grid applications using co-simulation. In: 2015 24th international conference on computer communication and networks (ICCCN). IEEE, pp 1–8
- Noori H (2012) Realistic urban traffic simulation as vehicular ad-hoc network (vanet) via veins framework. In: 2012 12th conference of open innovations association (FRUCT). IEEE, pp 1–7
- Stankovic JA (2014) Research directions for the internet of things. *IEEE Internet Things J* 1(1):3–9. <https://doi.org/10.1109/JIOT.2014.2312291>
- Wu S, Rendall JB, Smith MJ, Zhu S, Xu J, Wang H, Yang Q, Qin P (2017) Survey on prediction algorithms in smart homes. *IEEE Internet Things J* 4(3):636–644
- Wu D, Ren A, Zhang W, Fan F, Liu P, Fu X, Terpeny J (2018) Cybersecurity for digital manufacturing. *J Manuf Syst* 48:3–12
- Xu G, Yu W, Griffith D, Golmie N, Moulema P (2017a) Toward integrating distributed energy resources and storage devices in smart grid. *IEEE internet Things J* 4(1):192–204
- Xu H, Lin J, Yu W (2017b) Smart transportation systems: architecture, enabling technologies, and open issues. In: Secure and trustworthy transportation cyber-physical systems. Springer, Singapore, pp 23–49
- Xu H, Yu W, Griffith D, Golmie N (2018a) A survey on industrial internet of things: a cyber-physical systems perspective. *IEEE Access* 6:78238–78259
- Xu J, Guo H, Wu S (2018b) Indoor multi-sensory self-supervised autonomous mobile robotic navigation. In: 2018 IEEE international conference on industrial internet (ICII), pp 119–128. <https://doi.org/10.1109/ICII.2018.00021>
- Yang Q, Yang J, Yu W, An D, Zhang N, Zhao W (2014) On false data-injection attacks against power system state estimation: modeling and countermeasures. *IEEE Trans Parallel Distrib Syst* 25(3):717–729
- Yang X, Zhang X, Lin J, Yu W, Zhao P (2016) A gaussian-mixture model based detection scheme against data integrity attacks in the smart grid. In: 2016 25th international conference on, computer communication and networks (ICCCN). IEEE, pp 1–9

- Yang Q, An D, Min R, Yu W, Yang X, Zhao W (2017) On optimal PMU placement-based defense against data integrity attacks in smart grid. *IEEE Trans Inf Forensics Secur* 12(7):1735–1750. <https://doi.org/10.1109/TIFS.2017.2686367>
- Yu W (2012) False data injection attacks in smart grid: challenges and solutions. In: *Proceeding of NIST cyber security for cyber-physical system (CPS) workshop*
- Yu W, Griffith D, Ge L, Bhattarai S, Golmie N (2015) An integrated detection system against false data injection attacks in the smart grid. *Secur Commun Netw* 8(2):91–109
- Yu W, Liang F, He X, Hatcher WG, Lu C, Lin J, Yang X (2018) A survey on the edge computing for the internet of things. *IEEE Access* 6:6900–6919. <https://doi.org/10.1109/ACCESS.2017.2778504>

Internet Video Streaming

- [QoE Measurement and Assessment of Video Streaming](#)

Internet.org

- [Game Theoretic Modeling of Zero-Rating Internet Provisioning](#)

Inter-symbol Interference

- [Equalization Techniques for Single-Carrier Modulations](#)

Inter-vehicle Communication

- [Vehicular Ad Hoc Network](#)

Inter-vehicle Communication System

- [Vehicle Ad-Hoc Networks: Services, Security, and Privacy](#)

Intra-flow Network Coding

- [Joint Network Coding and Opportunistic Forwarding](#)

In-Vehicle Security

- [Automotive Security](#)

IoT

- [New Variants of Mirai and Analysis](#)

IoT Forensic Science

- [Internet of Things \(IoT\) Forensic Science](#)

IoT Forensics

- [Internet of Things \(IoT\) Forensic Science](#)

IoT Security

Lehlogonolo P. I. Ledwaba and
Gerhard P. Hancke
Department of Computer Science, City
University of Hong Kong, Kowloon, Hong Kong

Synonyms

[Cyber-physical system security](#); [Pervasive computing security](#)

Definitions

The Internet of Things is a large network of connected devices, which includes devices that are normally not capable of Internet connectivity or intelligent computation. The field of IoT security comprises all the different methods and countermeasures required to protect devices and networks within the Internet of Things with the goal of ensuring safe and reliable operations of the overarching systems and applications.

Historical Background

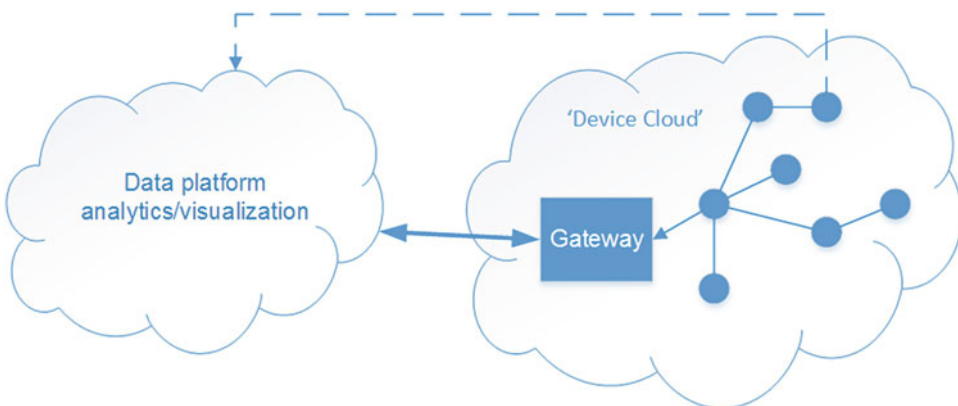
The Internet of Things has its roots in the earlier concepts of pervasive and ubiquitous computing. Everyday “dumb” devices could be equipped with embedded electronics to make them observe and interact with their surroundings. Following on from the progress of wireless sensor networks and automatic identification and data capture systems, the next step was to extend connectivity to these devices. This gave rise to the possibility of large networks of nontraditional computational devices. This leads to the generic architecture in Fig. 1, where we have a cloud of heterogeneous devices interacting with their surroundings and each other, which is connected to a back-end system, either directly via machine-to-machine communication or through a gateway. The back-end is responsible for processing potentially high-

volume data streams, data storage, analytics, and visualization. IoT architecture underpins many consumer and nonconsumer applications, from smart homes and healthcare to cyber-physical systems for industrial use in manufacturing and critical infrastructure monitoring.

As such, IoT devices are potentially an attractive target for attacks, and there has been several high-profile security incidents related to IoT devices. It is therefore critical that devices and networks are protected, but this is challenging. Securing devices is made difficult by a variety of device hardware and system restrictions including limited device energy, memory and processing resources, message size, and real-time operation. The scale in which IoT deployments are usually conducted requires that future solutions developed for the IoT should be highly scalable with an extended operational period. The IoT network should also be able to meet strict real-time deadlines, which means that communication latency and algorithm execution time should be predictable and minimal (Zhou et al. 2018).

Foundations

The subject of IoT security is very broad, covering security aspects of the devices, network communication, and the back-end platform. There are also many industry and academic opinions on this subject, with no definitive answer which aspects



IoT Security, Fig. 1 Generic system architecture

are most important. However, several high-level security themes have emerged that is found across different IoT-relevant security initiatives, such as the security frameworks by the Industrial IoT Consortium (ICC 2016) and OpenFog Consortium (OC 2016).

Tamper resistance: IoT devices are not deployed in physically secure locations and are open to tampering attacks. Attackers with access to devices can attempt several types of attacks to compromise device security, such as invasive, noninvasive, fault injection, and software attacks (Nisarga and Peeters 2016). Invasive attacks involve physical intrusion at device level, e.g., breaching the device enclosure to access device interfaces inside, or at integrated circuit (IC) level, e.g., depackaging an IC to access its internal structure. Noninvasive attacks observe external measurable behavior of the device, as security operations are carried out, and aim to extract sensitive data via side channels, e.g., electromagnetic analysis, power analysis, and timing analysis. Fault injection attacks occur when the attacker alters the environment or operating conditions of the device to cause a fault that compromises device security, e.g., over- or undervoltage attacks, over- or under-temperature attacks, and clock manipulation attacks. Software attacks can be launched by exploiting external debug interfaces, programming interfaces, and communication interfaces, which include the API security vulnerabilities, i.e., calling device commands/APIs, in ways that compromise security. To physically safeguard devices, the use of tamper-resistant hardware components and secure devices should be employed (ICC 2016). Mechanisms to detect tampering may be built into the device casing, such as enclosure monitoring sensors such as light, temperature, pressure, or vibration sensors (Nisarga and Peeters 2016), and attempts to access device hardware should be reported and devices in question isolated from the network. Tamper mesh can be employed for active tamper monitoring of the IC packaging but is typically only employed on ICs that are marketed as secure or security-hardened. Electromagnetic leakage shields could also be incorporated within the device enclosure

or IC packaging. The addition of amplitude or temporal noise in hardware or software could decrease the signal-to-noise ratios of side channels to counteract attacks.

Trusted execution: In addition to tamper resistance, the establishment of a root of trust is crucial for supporting mechanisms for device identification and integrity checking (ICC 2016). To create a secure root of trust for a networked device, there should ideally be a hardware root-of-trust mechanism such as a hardware-based trusted execution environment (TEE) implemented with a trusted platform module (TPM) (TCG 2008). TPMs can be used to provide storage, processing, and cryptographic functions such as encryption, decryption, key generation and storage, and data signing in a secured and isolated execution environment. This offers physical tamper resistance to protect sensitive data and cryptographic processing capture and physically and logically separates security and cryptographic functions from normal device operations. It can be used as the basis of secure device management and software provisioning, e.g., authenticity of the new applications verified, and the attestation of device integrity. Newer embedded processor architecture is starting to incorporate integrated, system-on-chip TPMs, while older devices could be hardened by adding an external TPM, such as a smartcard-based secure access module (SAM).

Attestation: The basis of attestation is that “the entity that is to be tested, called the prover, sends a status report of its current configuration to another party, called the verifier, to demonstrate that it is in a known and thus trustworthy state” (Asokan et al. 2015; Valente et al. 2014). This technique is utilized toward the assurance of device integrity. Remote attestation schemes assume that the prover is provided with a trusted mechanism such as a TPM, with integrity measurements being taken and securely stored during the secure boot process. When conducting the attestation, the verifier sends a request for the device configuration measurements, and the prover retrieves, hashes and signs the measurements through the use of a digital signature algorithm, and then sends this to the verifier. The

verifier then verifies the signature and compares the measurements against the expected measurements for the device configuration (Valente et al. 2014). Given the potentially large-scale and heterogeneous nature of IoT deployments, attestation is a challenging problem as it is difficult for the verifier to keep track of every possible device configuration connected at a particular time to the network. Devices might also not have a TPM and need to relay on software-based attestation, and ideally device and state integrity measurements should also be verifiable after boot to achieve run-time attestation (Asokan et al. 2015; Ibrahim et al. 2016).

Cryptographic mechanisms: Endpoint devices use cryptographic mechanisms to maintain confidentiality and integrity of execution state and data. Cryptographic mechanisms also secure the communications exchanged within the IoT network by providing authenticated, confidential channels between devices, gateways, and back-end systems. There are many well-known algorithms to choose from for building security mechanisms for symmetric encryption, e.g., AES; asymmetric encryption and digital signature, e.g., RSA; and data hashing, e.g., SHA-2. Elliptic-curve cryptography also allows for computationally efficient asymmetric signing, e.g., ECDSA, and key agreement, e.g., ECDH, with smaller key sizes compared to conventional asymmetric algorithms. Performing cryptographic operations on IoT endpoint devices has been considered a challenge owing to their resource-constrained nature. There is still merit in finding more efficient cryptographic algorithm implementations for IoT devices as in some applications memory and operational latency is important. However, with recent advances in processing and storage technologies, new generation of IoT processors is feasibly capable of running standard symmetric and asymmetric cryptographic without negatively impacting the device ability to support other network operations (Ledwaba et al. 2018). This allows the possibility of using and optimizing standard, strong cryptographic functions rather than there

being a need to design new cryptographic algorithms.

System monitoring and intrusion detection: Within the IoT application space, sensor devices capture and transmit data pertaining to their deployment environment up toward the main network processing environment for analysis and actuation in data streams (Bertino et al. 2016). Owing to the nature of IoT networks, infiltration, data capture, and modification may occur without detection as streams are transferred from the network edge up to the cloud. To assist in providing the continued integrity in the network operations, intrusion detection and prevention techniques are often utilized. In comparing historic analytics of device data and network traffic with real-time data, these intrusion detection techniques are then capable of signaling to the wider network when unauthorized device or network behavior occurs (Midi et al. 2017). This aids in detecting security attacks on the system early in the process, making a successful defense more likely. The main problem is that while data analytic methods are moving forward, the provision of data verification, confidentiality, and integrity of the data streams generated in an IoT network remains an open research question, while device monitoring solutions for the IoT such as intrusion detection are still largely academic.

Key Applications

The Internet-of-Things paradigm is used in a number of applications ranging from consumer electronics to industrial and cyber-physical systems. IoT security plays a role in protecting such systems and ensuring their safe and reliable operation.

Cross-References

- [Anomaly Detection for IoT Systems](#)
- [Wireless Sensor Network Security](#)

References

- Asokan N, Brasser F, Ibrahim A, Sadeghi A, Schunter M, Tsudik G, Wachsmann C (2015) SEDA: scalable embedded device attestation. In: ACM SIGSAC conference on computer and communications security, pp 964–975
- Bertino E, Choo K, Georgakopoulos G, Nepal S (2016) Internet of things (IoT): smart and secure service delivery. *ACM Trans Internet Technol* 16:221–227
- Ibrahim A, Sadeghi A, Tsudik G, Zeitouni A (2016) DARPA: device attestation resilient to physical attacks. In: ACM conference on security and privacy in wireless and mobile networks, pp 171–182
- ICC (2016) Industrial Internet Consortium: industrial internet of things volume G4: security framework. http://www.iiconsortium.org/pdf/IIC_PUB_G4_V1.00_PB-3.pdf
- Ledwaba L, Hancke G, Venter H, Isaac S (2018) Performance costs of software cryptography in securing new-generation internet of energy endpoint devices. *IEEE Access* 6:9303–9323
- Midi D, Rullo A, Mudgerikar A, Bertino E (2017) Kalis: a system for knowledge-driven adaptable intrusion detection for the internet of things. In: IEEE international conference on distributed computing systems, pp 656–666
- Nisarga B, Peeters E (2016) System-level tamper protection using MSP MCUs. <http://www.ti.com/lit/an/slaa715/slaa715.pdf>
- OC (2016) OpenFog Consortium: OpenFog reference architecture for Fog computing. https://www.openfogconsortium.org/wp-content/uploads/OpenFog_Reference_Architecture_2_09_17-FINAL.pdf
- TCG (2008) Trusted Computing Group: trusted platform module (TPM) summary. https://trustedcomputinggroup.org/wp-content/uploads/Trusted-Platform-Module-Summary_04292008.pdf
- Valente J, Barreto C, Cardenas A (2014) Cyber-physical systems attestation. In: IEEE international conference on distributed computing in sensor systems, pp 354–357
- Zhou L, Yeh KH, Hancke G, Liu Z, Su C (2018) Implementing security and privacy for the industrial internet of things. *IEEE Signal Process Mag* 965: 325–329

IP Mobility

- [Proxy Mobile IPv6](#)

IP Mobility Management

- [Network-Layer Mobility Management](#)

Jamming

- [Interference Mitigation in Wireless Communications-Based Train Control \(CBTC\), Cognitive Control Approach](#)

Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks

Yiming Liu¹, Richard Fei Yu^{2,3}, Xi Li¹, Hong Ji¹, Heli Zhang¹, and Victor C. M. Leung⁴

¹School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, China

²Carleton University, Ottawa, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

⁴The University of British Columbia, Vancouver, BC, Canada

Synonyms

[Adaptive video streaming](#); [Content delivery](#); [Mobile edge computing](#); [Wireless edge caching](#)

Definitions

Adaptive video streaming is a popular video delivery technique which could adjust the quality level of a video stream to adapt the dynamic network conditions and support heterogeneous devices (e.g., TV, PC, tablet, and smartphones). By utilizing video transcoding, the original content could be transcoded to a preconfigured format (e.g., codec, resolution, and bitrate) for meeting the requirements of users. In general, it is necessary to manage resources to cache video contents and transcode them to overcome the diversity issues presented by the multiple formats of video streams. However, the on-demand provisioning, scalable resource management, and routing optimization pose significant challenges to the existing network operators.

Wireless edge caching, as an extension of in-network caching, can efficiently reduce the duplicate content transmission and improve the quality of services (QoS) of users. By caching a number of popular contents proactively at the network edge, the desired contents can be delivered with the short-range communications between the small base stations (SBSs) and users directly. This is able to significantly improve the network performance, e.g., transmission delay and energy efficiency. However, due to the limited storage size of network edge, it is important to decide

what content to cache and how to select route-path taking into account content popularity and network conditions.

Mobile edge computing (MEC), which utilizes the powerful computing capabilities at the edge of the networks, has been recognized as a promising solution to process and transcode video streams for the users. The basic idea behind MEC is that by transcoding video into different formats in close-proximity to the users, huge backhaul burden is released and user's experience is improved. When the SBSs do not have enough computing resources to process their video contents, they have to perform routing selection and forward their users' requests to the remote cloud server at the expense of consuming more network resources and higher communications latency. Thus, the resource management and routing optimization issues play an important role in improving resource utilization and QoS of users.

Historical Background

Traditionally, video streaming solutions utilize either a UDP-based (RTP/RTSP and IP-layer multicast) or TCP-based (RTMP) approach to optimize video delivery efficiency. Currently, the HTTP-based solution is widely used due to the advantages of its extensive infrastructure, servers, and caching. However, in varying wireless networks, it is critical to adapt to the network conditions for providing better services to users. To overcome such challenges, in adaptive bitrate streaming (HLS (HTTP Live Streaming (HLS) 2014), DASH (Stockhammer 2011)), the video file is partitioned into multiple segments each of which has a duration of a few seconds. Then the service server maintains different bitrates of the same video segments and delivers them to users with distinct preferences. Recently, edge caching and computing techniques attract a lot of researches for improving the performance and efficiency of video distribution, by carefully allocating network resources (e.g., caching, transcoding, and routing).

Wireless edge caching refers to that SBSs cache the video segments and delivers them to the users directly, reducing response delay and

network traffic. The benefits of wireless edge caching have been demonstrated in various types of wireless networks (Tan et al. 2018; Wang et al. 2017). Due to the limited caching resources and large-scale contents within networks, it is beneficial to apply popularity prediction (Cho et al. 2012) and proactive caching (Zhang et al. 2017) for deciding what content to cache. On the other hand, routing policies, deciding where to obtain the contents, are also crucial for overall caching performance. Deh et al. (Liu et al. 2017) propose a unified optimization for the joint caching and multi-relay problem to improve the caching and bandwidth utilization. He et al. (He and Song 2016) propose a hop-by-hop routing protocol, which implements the optimization solutions by generating a set of flow-splitting and routing decisions to reduce the overall networking cost.

MEC has recently received significant research attention for handling video services. A MEC-based architecture, where each MEC server could modify the bitrate version of the video on the fly according to current network conditions, is proposed in (Liang et al. 2017) to improve the performance of adaptive video streaming. Considering the varying wireless channel conditions, a video bitrate adaptation algorithm is proposed in Pedersen and Dey (2016) to joint optimize radio access network caching and computing. To improve resources utilization, the authors in Liu et al. (2018) investigate the dynamic resource allocation for transcoding and delivering adaptive video streaming services. By combining caching and routing issues, the authors in Jin et al. (2015) investigate the NFV-based resource allocation and SDN-based routing for cost-efficient video distribution over future networks.

Foundations

In wireless networks, a large number of end devices, such as smart devices, wearable devices, tablet, connect to the SBSs through reliable and efficient access networks. These end devices may generate amounts of video service and application requests that are sent to the network edge. Then, the SBSs with the capabilities of

caching, computing, and network connectivity will process and transmit video contents to different users. Because of the tremendous amount of accessible contents over the Internet and limited storage at the SBSs, those caching contents cannot include all of them within the networks. Since each SBS caches partial content segments, there would be three possible cases to provide content to the user, including exact cache case, transcoding cache case, and cache miss case.

The content response delay, which can be directly translated into user's QoS, plays an important role in improving the user's experience. In the case of exact cache scenario, since the SBS transmits the content segments directly to the user, there is only access delay, which can be derived by following the rule that the ratio of the total bitrate of the required segments and the transmission rate. In the case of transcoding cache, the SBS has to transcode the higher bitrate version into the required bitrate version, which may incur transcoding delay and computing resource cost. Then, the delay of the transcoding cache case can be derived as the sum of the access delay and the transcoding delay. Lastly, in the case of cache miss, the SBS has to forward the request to the remote cloud server and obtains the content segment via the backhaul links. It will incur great delay and backhaul cost. The total cache miss delay includes the access delay and the backhaul delay that are related to the available backhaul rate.

To access a piece of content, a user must decide whether to route its request to an SBS or to the remote cloud server. Resource management policy (e.g., caching and computing resources) should also be well designed to minimize the content response delay. To address these issues, the optimization problem is divided into two stages: routing selection and resource allocation. Firstly, when the user broadcasts the content request to the neighboring SBSs, each SBS that receives the request will estimate the delay by providing all the required contents to the user with a given resource allocation policy. Under the resource constraints, each user would select a set of SBSs following the rule that the total delay is minimized. Then, given the routing selection, caching and computing resources allocation

problem could be formulated as a convex optimization problem and is easy to solve.

Key Applications

The joint caching, computing, and routing strategies play an important role in adaptive video streaming services. It can be further applied to various types of systems, for instance, intelligent medical system, monitoring system, and vehicular networks.

Cross-References

- [Integrated System of Networking, Caching, and Computing](#)
- [Mining Task Offloading in Wireless Blockchain Networks](#)
- [Mobile Content Distribution Architecture](#)

References

- Cho K, Lee M, Park K, Kwon TT, Choi Y, Pack S (2012) Wave: popularity-based and collaborative in-network caching for content-oriented networks. In: Proceedings of the IEEE INFOCOM WKSHPS, pp 316–321
- He J, Song W (2016) Optimizing video request routing in mobile networks with built-in content caching. *IEEE Trans Mob Comput* 15:1714–1727
- HTTP Live Streaming (HLS). IETF internet-drafts 2014. <https://tools.ietf.org/html/draft-pantos-http-live-streaming-13/>
- Jiu Y, Wen Y, Westphal C (2015) Towards joint resource allocation and routing to optimize video distribution over future internet. In: Proceedings of the IEEE IFIP networking, pp 1–9.
- Liang C, He Y, Yu FR, Zhao N (2017) Video rate adaptation and traffic engineering in mobile edge computing and caching-enabled wireless networks. In: Proceedings of the IEEE VTC-Fall, pp 1–5
- Liu B, Zhang H, Ji H, Li X (2017) A novel joint transmission and caching optimizing scheme in multirelay networks: video service quality assurance scheme. *Int J Commun Syst* 30(14):e3284
- Liu Y, Yu FR, Li X, Ji H, Zhang H, Leung VC (2018) Joint access and resource management for delay-sensitive transcoding in ultra-dense networks with mobile edge computing. In: Proceedings of the IEEE ICC, pp 1–6
- Pedersen HA, Dey S (2016) Enhancing mobile video capacity and quality using rate adaptation, RAN

caching and processing. *IEEE/ACM Trans Netw* 24:996–1010

Stockhammer T (2011) Dynamic adaptive streaming over HTTP—standards and design principles. In: *Proceedings of the 2nd ACM conference on multimedia systems*, pp 133–144

Tan Z, Yu FR, Li X, Ji H, Leung VC (2018) Virtual resource allocation for heterogeneous services in full Duplex-enabled SCNs with mobile edge computing and caching. *IEEE Trans Veh Technol* 67:1794–1808

Wang C, He Y, Yu FR, Chen Q, Tang L (2017) Integration of networking, caching and computing in wireless systems: a survey, some research issues and challenges. *IEEE Commun Surv Tutor* 20(1):7–38

Zhang H, Liu B, Li M, Ji H, Li X, Leung VC (2017) Pricing, caching selection, and content delivery in wireless networks: a hierarchical approach. In: *Proceedings of the IEEE GLOBECOM*, pp 1–6

Joint Laplace Transform of Co-channel Interference

► [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)

Joint Network Coding and Opportunistic Forwarding

Chen Zhang^{1,2,3}, Yuanzhu Chen¹, and Somayeh Kafaie⁴

¹Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

²The George Washington University, Washington, DC, USA

³Southeast University, Nanjing, China

⁴Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Canada

Synonyms

[Inter-flow network coding](#); [Intra-flow network coding](#); [Opportunistic routing](#); [Secure network coding](#)

Definition

Opportunistic routing and network coding are promising network paradigms. Both of them benefit from the broadcast nature of wireless media and will significantly improve network performance. Network coding applies algebraic algorithms to code data packets; each coded packet randomly combines partial information of different data packets, which extends the distribution of data packets in the network. Since each coded packet is equally beneficial and inherently different and each coded packet contains part of information from different original packets, no coded packet is special or indispensable. Thus, even though some coded packets are lost, the sender can keep forwarding coded packets without learning which packets are lost. Opportunistic routing considers multiple downstream nodes are potential next-hop forwarders, which can effectively reduce the possibilities of unnecessarily reconstruction path or retransmission due to link breakage on a preselected path.

Historical Background

Network coding represents an innovative idea introduced by Ahlswede et al. in 2000 (Ahlswede et al. 2000). It has emerged as an important potential approach to the operation of communication networks, especially wireless networks. The major benefit of network coding comes from its ability to code data, across time and across flows. It makes data transmission over lossy wireless networks robust and effective. There has been a rapid growth in the theory and potential applications of network coding. These research can be summarized in two types of network coding: inter-flow network coding that mixes packets from different sources and intra-flow network coding that is about fusing packets from the same source.

Opportunistic routing allows multiple nodes to overhear packet transmissions and be involved in further forwarding packets. With intra-flow network coding, packets from the same flow are coded; each coded packet is equally bene-

ficial and inherently different. Since no special coded packet is dispensable, the deficit information from lost packets can be easily filled by following coded packets. The combination of these two ideas in a natural fashion to provide opportunistic routing without a coordination and exploit opportunity gains inherent in wireless media.

Foundations

In general, the linear network coding is different from the *xor* operation by a linear combination of the data, which interprets data over some finite fields, e.g., $GF(2^s)$. It allows a much larger degree of flexibility to combine packets. With the linear network coding, coded packets are linear combinations of original packets. The coding process of addition and multiplication is performed over the finite field. The finite field or Galois field is a field that contains a finite number of elements. A finite field is a set on which the operation of multiplication, addition, subtraction, and division are defined and satisfy certain basic rules (Lidl and Niederreiter 1997). Addition is the standard bitwise *xor*. Division is computed by the Euclidian algorithm (Fragouli et al. 2006a). Both multiplication and division can be implemented efficiently with s shifts and additions (Wagner 2003).

Intra-flow network coding will code packets from the same flow. One simple example of the intra-flow network coding can consider a linear topology with three nodes. Source node S will directly send n packets to destination D . The packet delivery probability is 50% in both directions. Without network coding, the source node requires the ack to learn whether a packet is received by the destination. Thus, the expected number of transmissions will be $2n + 2n$ ($2n$ for data packets, $2n$ for acks). Network coding allows nodes to make the random linear combination of data packets. Once the destination receives n independent coded packets, it can decode n original packets. Then the destination sends one ack for all n packets. The expected number of transmissions will be $2n + 2$.

The detail discussion about intra-flow network coding as follows.

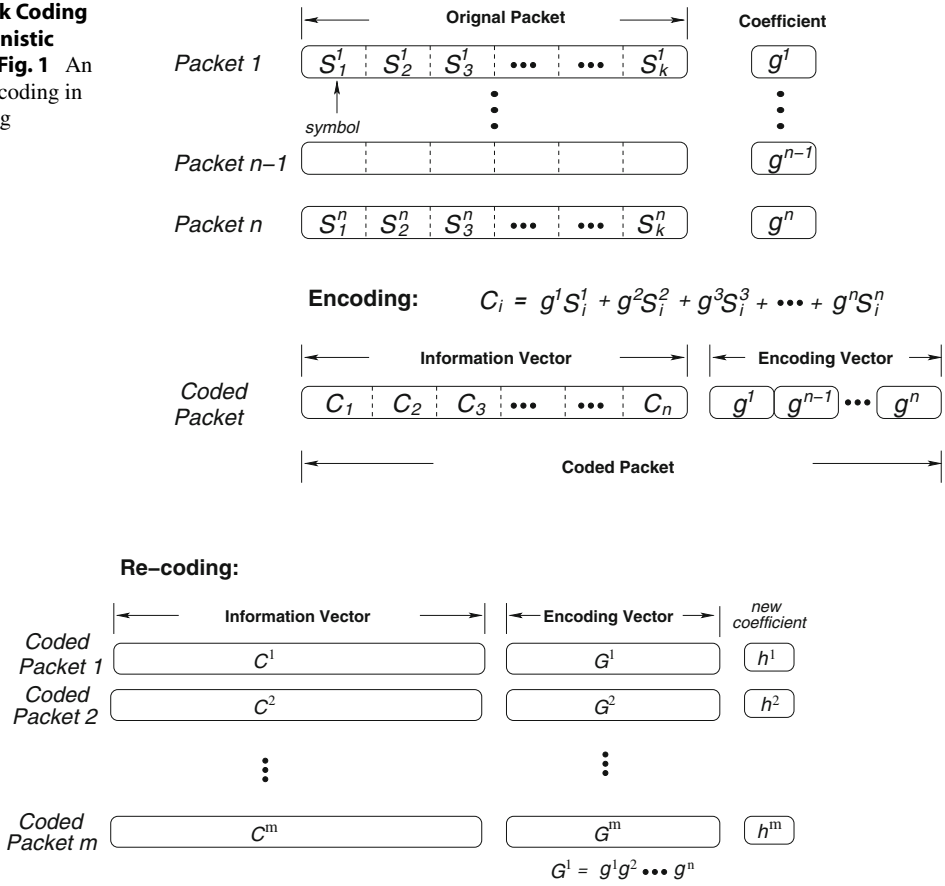
Encoding at Source

Assume that the number of original packets is defined as $O^1, O^2, \dots, O^n$ and each packet consists of L bits. Original packets will be divided into several symbols with the same length. When original packets to be combined do not have the same size or the length of the original packets cannot be equally divided into symbols with the same length, the shorter packets are padded with trailing 0 s. The s consecutive bits of a packets are defined as a symbol over the finite field $GF(2^s)$; therefore, each packet consists of $K = L/s$ symbols. In linear network coding, each packet through the network is associated with a sequence of coefficients $g^1, g^2, \dots, g^n$ in $GF(2^s)$. The coded packet is equal to $C = \sum_{i=1}^n g^i O^i$. The summation will occur at every symbol position k ; for example, the coded symbol $C_k = \sum_{i=1}^n g^i S_k^i$, where S_k and C_k are the k th position of original symbol and coded symbol, respectively. We show one example of the encoding process in Fig. 1.

After packets are coded together, each coded packet contains both a set of coefficients $G = \{g^1, g^2, \dots, g^n\}$, called encoding vector and coded data $C = \sum_{i=1}^n g^i O^i$, called information vector (Fragouli et al. 2006a). The encoding vector is used by receivers to decode original packets. For example, the encoding vector $G = \{0, \dots, 1, \dots, 0\}$, where the 1 is at the i th position. It means that the information vector is equal to the original packet O^i .

In the linear network coding, the selection of the coefficients for each linear combination is an important issue for practical considerations. A simple algorithm is that each node in the network selects random coefficients over the field $GF(2^s)$ in a completely independent and decentralized manner (Ho et al. 2003). With the random network coding, there is a still certain probability of selecting linearly dependent combinations. However, this probability can be reduced by choosing large field size 2^s (Fragouli et al. 2006a). Simulation results show that even for small field

Joint Network Coding and Opportunistic Forwarding, Fig. 1 An example of encoding in network coding



Joint Network Coding and Opportunistic Forwarding, Fig. 2 An example of recoded in network coding

sizes (e.g., $s = 8$), this probability is still negligible (Wu et al. 2004).

Recoding at Intermediate Nodes

The encoding process in linear network coding can be performed recursively. Therefore, the intermediate node can recode the coded packet. Consider that a node receives a set of coded packets $(G^1, C^1), \dots, (G^m, C^m)$, shown in Fig. 2, where G^m is the encoding vector and C^m is the information vector. This intermediate node will generate a new set of coefficients $H = \{h^1, h^2, \dots, h^m\}$ and compute the linear combination $RC = \sum_{i=1}^m h^i C^i$. The new corresponding encoding vector G' is not simply equal to H , since the coefficients need to be generated with respect to the original packet $O^1, O^2, \dots, O^n$, the

new encoding vector will be $G' = \sum_{i=1}^m h^i G^i$ (Fragouli et al. 2006a). This operation can be repeated at intermediate nodes several times in the network.

Decoding at Destination

Assume that the destination node receives a set of coded packets $(G^1, C^1), \dots, (G^m, C^m)$. In order to decode the original packets, the destination node needs to solve the linear system $C^i = \sum_{j=1}^n G^i O^j$, where $G^i = g_i^1 \dots g_i^n$

$$\begin{pmatrix} C^1 \\ \vdots \\ C^m \end{pmatrix} = \begin{pmatrix} g_1^1 \dots g_1^n \\ \vdots \\ g_m^1 \dots g_m^n \end{pmatrix} \times \begin{pmatrix} O^1 \\ \vdots \\ O^n \end{pmatrix} \quad (1)$$

The unknowns are O^i , and this system is a linear system with m equations and n unknowns. If the number of received packets in the destination node is equal to or larger than the number of unknowns, the destination has a chance of decoding all original packets. Here, the decoding condition that the number of received packets is larger than the number of unknowns ($m < n$) is not sufficient, because some of the coded packets may be linearly dependent.

The destination node will solve a set of linear equations. This process can be done as follows. It stores the encoded vectors from the received coded packets in a decoding matrix (Fragouli et al. 2006a). When an encoded packet is received, the encoded vector of this packet will be inserted as the last row of the decoding matrix. When this matrix will be transformed based on the current encoded vectors, the matrix will be transformed to a reduced row echelon form using Gaussian elimination, where the leading term of each row is 1 and the position of the leading terms moves to the right. A received packet is called innovative if its encoded vector increases the rank of the matrix. As soon as the matrix contains n rows, where n is the number of original packets, the destination node can decode all original packets. The size of the decoding matrix has great impact on the decoding delay at the destination (Chou et al. 2003). For practical considerations, this size of decoding matrix should be limited. Therefore, packets are usually grouped into batches or generations, and only packets of the same generation can be combined. The size of batches also has significant impact on the performance of network coding (Fragouli et al. 2006b).

Advantage of Jointed Protocols

Opportunistic routing improves network performance by exploiting the broadcast nature of wireless media. The main concern of opportunistic routing is how to utilize all possible paths to deliver packets with available cost. A reliable coordination mechanism for opportunistic routing is always challenge work and hard to carry out accurately. The protocols of using network cod-

ing can eliminate the unnecessary acknowledgment in opportunistic routing. The architecture design of network layers can be independent since no feedback information is required from MAC layer to network layer. On the other hand, the major challenge of network coding is how to find an efficient way to explore the coding opportunistic in the network for which opportunistic routing seems a promising approach. Therefore, the advantage of joint protocols is to use the strengths of one to overcome the weakness of the other (Ho et al. 2003).

Key Applications

Network coding applications deal with unreliable transmissions in wireless and content distribution networks. Opportunistic routing can further explore the coding opportunity. They are expected to be critical technologies for networks of the future. Multiple unicasts, security, networks with unreliable links, and quantum networks can be discussed with network coding.

Cross-References

- Delay-Tolerant Network Routing
- Distributed Storage Security
- Opportunistic Routing

References

- Ahlswede R, Cai N, Li S-YR, Yeung RW (2000) Network information flow. *IEEE Trans Inf Theory* 46(4): 1204–1216
- Chou PA, Wu Y, Jain K, Chou PA (2003) Practical network coding. In: Proceedings of allerton conference on communication, control, and computing. [Online]. Available: <https://www.microsoft.com/enus/research/publication/practical-network-coding-2>
- Fragouli C, Boudec J-YL, Widmer J (2006a) Network coding: an instant primer. *ACM SIGCOMM Comput Commun Rev* 36(1):63–68
- Fragouli C, Widmer J, Boudec J-YL (2006b) A network coding approach to energy efficient broadcasting: from theory to practice. In: Proceedings of IEEE international conference on computer communications, Barcelona, pp 1–11

- Ho T, Koetter R, Medard M, Karger DR, Effros M (2003) The benefits of coding over routing in a randomized setting. In: Proceedings of IEEE international symposium on information theory, Yokohama, pp 442–442
- Lidl R, Niederreiter H (1997) Finite fields, vol 20. Cambridge University Press
- Wagner NR (2003) The laws of cryptography with java code. Available online at Neal Wagner's home page
- Wu Y, Chou PA, Jain K (2004) A comparison of network coding and tree packing. In: Proceedings of IEEE inter-

national symposium on information theory, Chicago, pp 145–146

Joint Precoding/Processing

- [Base Station Cooperations Under Imperfect Conditions](#)

Key 4G/LTE Technologies

► [Key Technologies in 4G/LTE Network](#)

Key Technologies in 4G/LTE Network

Chengyu Wu and Weiqiang Xu
Zhejiang Sci-Tech University, Hangzhou, China

Synonyms

[Key 4G/LTE technologies](#); [Wireless 4G/LTE key technologies](#)

Definition

Long-Term Evolution (LTE) is a standard developed by the 3GPP (3rd Generation Partnership Project) for high-speed mobile communication based on the GSM/EDGE and UMTS/HSPA technologies, commonly known as 3.9G (beyond 3G but pre-4G), that is, the version from HSPA transition to 4G. LTE is specified in the 3GPP Release 8 and 9 document series, and it can provide 300 Mbit/s peak rates for downlink, 75 Mbit/s for uplink, and less than 5 ms for latency. 4G

is a continual improvement standard to the LTE radio technology and architecture, which is also called International Mobile Telecommunications Advanced (IMT-Advanced) as defined by the International Telecommunications Union-Radio communications sector (ITU-R) (https://en.wikipedia.org/wiki/LTE_Advanced). 4G is specified in the 3GPP Release 10 and known as LTE-Advanced, and it can provide 100 Mbit/s peak rates for high mobility and approximately 1 Gbit/s for low mobility (<https://en.wikipedia.org/wiki/4G>). On the whole, LTE is pre-4G technology that brings the 3G networks to a new 4G performance era radically; LTE-Advanced is real 4G technology that reaches the target set for IMT-advanced (ITU-R 2008).

To enable better performance than 3G, 4G/LTE network has applied some efficient key technologies (3GPP TR 36.814 2010): (1) Orthogonal Frequency Division Multiplexing (OFDM); (2) Multiple-input Multiple-output (MIMO); (3) Mobile IP; (4) Software Defined Radio (SDR); (5) Link adaptation, and (6) Small cell.

Historical Background

Currently, 4G/LTE is good enough for most mobile users and will be there for some time (<https://en.wikipedia.org/wiki/4G>). To understand 4G/LTE, it is essential to look back to the ongoing process of wireless cellular technologies since the incipency of 1G almost

forty years ago. 1G was introduced in the early 1980s, which is an analog communication system that established seamless mobile connectivity introducing mobile voice services at speeds of up to 2.4 kbit/s. 2G digital wireless technologies can increase voice capacity, and the most popular 2G technology was Global Systems for Mobile Communications (GSM) with SMS messaging. 3G, which is required to meet IMT-2000 technical standards, improved the data speeds up to 2 Mbit/s for indoor or stationary users, 384 kbit/s for pedestrians, and 144 kbit/s for moving vehicles. In more recent years, 3G has made some advancements, for example, HSPA, which is often called 3.5G or Turbo 3G, can offer theoretical speeds up to 7.2 Mbit/s. Better than 3G, 4G/LTE has many specific characteristics (Anitha and Joshphin 2015).

- (a) **Faster Speed:** As we mentioned above, the 4G/LTE system should provide more than 100 Mbit/s peak rates at rest and an average of 20 Mbit/s when in motion. It is clear that the speed is much higher than 3G.
- (b) **Seamless handover:** 4G/LTE networks support global roaming across multiple wireless and mobile networks. This provides greater flexibility for vendor selection.
- (c) **Better network capacity:** The capacity at least 10 times than that of 3G, which means that the download time of a 10 Mbyte file is quickened to one second on 4G/LTE networks comparing with 200 s on 3G networks. This enables better user experience.

Foundations

Key technologies in 4G/LTE networks are mainly described below:

Orthogonal Frequency Division Multiplexing (OFDM)

In OFDM, the signal is divided into orthogonal subcarriers, the signal is “narrowband” on each subcarrier and therefore is not affected by multipath, as long as a guard interval is inserted between each OFDM symbol. OFDM can pro-

vide (1) significant advantages for physical layer performance and (2) a framework for improving layer 2 performance by providing additional degrees of freedom (Al-Kebsi et al. 2009). Using OFDM, the time domain, air domain, frequency domain, or even code domain can be utilized to optimize the use of wireless channels. In OFDM, the transmission in a multipath environment is very robust and the receiver complexity is lowered.

Orthogonal Frequency Division Multiple Access (OFDMA), a multiple access technique, is the evolution of OFDM technology. In this case, each OFDM symbol may use a different set of subcarriers (subchannels) to send information to multiple users or to send information from multiple users (Holma and Antti 2009). This can provide additional flexibility for resource allocation (increase capacity) and enable cross-layer optimization of spectrum usage.

The major disadvantage of an OFDM system is High Peak to Average Power Ratio (PAPR) due to the summation of a large number of subcarriers coherently. 4G/LTE system have applied several techniques to reduce the PAPR, including clipping, peak windowing, and peak cancellation.

Multiple-input Multiple-output (MIMO)

MIMO referred to the use of multiple antennas at the transmitter and the receiver to achieve the goal of 4G/LTE networks (Di Renzo et al. 2014). MIMO can improve the spectrum efficiency of 4G/LTE system directly or indirectly. MIMO technology can be classified into four main categories: (1) spatial diversity, (2) beamforming, (3) space division multiplexing, and (4) space division multiple access. Spatial diversity and beamforming can improve the link quality, enhance coverage, and indirectly improve the spectrum efficiency; space division multiplexing and space division multiple access can improve the user capacity and directly improve the spectrum efficiency. In spatial diversity, a single stream is transmitted using space-time coding, which exploits the independent fading in the multiple antenna links to enhance signal diversity. Beamforming is a signal precoding technology based on the antenna array, and it can produce

directivity beam by adjusting the weighting coefficient of each antenna array, thus to obtain obvious array gain. In space division multiplexing, a high-rate signal is split into multiple lower-rate streams and each stream is transmitted from a different transmit antenna in the same frequency channel for link capacity enhancement.

Mobile IP

Mobile IP (or MIP) is an Internet Engineering Task Force (IETF) standard communications protocol that is designed to allow mobile device users to move from one network to another while maintaining a permanent IP address (https://en.wikipedia.org/wiki/Mobile_IP). Data may be lost during the time when a mobile node moves between domains of a network and hence effects the Quality of Service (QoS) of the mobile node. In order to solve this problem, the mobility management scheme is used to reduce the movement detection latency and maintain the IP connectivity. On the other hand, Session Initiation Protocol (SIP) is used for signaling and controlling multimedia communication sessions in applications of Internet telephony for voice and video calls to support fast handoff (https://en.wikipedia.org/wiki/Session_Initiation_Protocol).

Software Defined Radio (SDR)

The final form of a 4G/LTE device constitutes various standards, such like 2G, 3G, 4G. SDR is a single radio that can be used for many modes by simply reconfiguring the radio with different software (Tripathi et al. 2015). Such a design produces a radio which can receive and transmit widely different radio protocols based solely on the software used (https://en.wikipedia.org/wiki/Software-defined_radio). This flexible architecture can improve performance in terms of power consumption and reduced interference. SDR becomes an enabler for the aggregation of multistandard pico-/microcells in 4G/LTE networks. This is a powerful aid to provide multistandard, multiband equipment through simultaneous multichannel processing.

Link Adaptation

The basic idea of link adaptation is to adjust transmission parameters and schemes to take advantage of variations in channel conditions (Zheng et al. 2005). The most efficient link adaptation techniques are adaptive modulation and coding (AMC) and hybrid-automatic request repeat (H-ARQ). AMC changes the modulation and coding format according to variations in channel conditions, through explicit signal-to-noise ratio (SNR) measurements. H-ARQ is a combination of high-rate forward error-correcting coding and ARQ error-control, and it is an implicit link adaptation technique (https://en.wikipedia.org/wiki/Hybrid_automatic_repeat_request). It is the best use of combining AMC with H-ARQ in 4G/LTE networks, while AMC provides gross data rate selection as well as that H-ARQ adjusts the data rate based on channel conditions. In 4G/LTE system with OFDM and MIMO technologies, link adaptation can improve rate of transmission, and/or bit error rates, and exhibit great performance enhancements compared to systems that do not exploit channel knowledge at the transmitter (https://en.wikipedia.org/wiki/Link_adaptation).

Small Cell

In 4G/LTE networks, the principle of heterogeneous network is introduced where the mobile network is constructed with layers of small and large cells. Small cells are low-powered cellular radio access points that operate in licensed or unlicensed spectrum. Compared to a mobile macrocell, they have a shorter range and typically handle fewer concurrent calls/sessions. In 4G/LTE system, larger numbers of small cells are deployed to increase cellular network capacity, quality, and resilience. Small cells may encompass femtocells, picocells, microcells, and distributed radio technology using centralized baseband units and remote radio heads. The cost advantages of small cells compared with macrocells make it economically feasible to provide coverage of much smaller communities. Small cells also help service providers discover new revenue opportunities through their location

and presence information (https://en.wikipedia.org/wiki/Small_cell).

Key Applications

Compared with previous mobile network technologies, 4G/LTE offers much higher bandwidth, lower latency, and improved spectrum efficiency (Akayama et al. 2012). In practice, this activates more applications to be used on mobile devices, improves the performance of many existing applications, and makes feasible new applications depending on reliable high-speed such as real-time sharing of large files and streaming media. The improved convenience, user experience, and practicality of 4G/LTE also hasten uptake of those existing applications that already work on mobile devices.

Cross-References

- [Heterogeneous Small Cell Networks](#)
- [MIMO](#)
- [Mobile IP](#)
- [OFDM](#)

References

- 3GPP TR 36.814 (2010) Further advancements for E-UTRA physical layer aspects (LTE-Advanced) (Release 9)
- Akayama S, Schlautmann A, Place J, Keeping S (2012) The business benefits of 4G LTE. Arthur D. Little. <http://www.adlittle.com/en/insights/viewpoints/business-benefits-4g-lte>
- Al-Kebsi II et al (2009) A novel algorithm with a new adaptive modulation form to improve the performance of OFDM for 4G systems. In: International conference on future computer and communication. IEEE, pp 11–15
- Anitha X, Joshphn J (2015) 4G mobile communications-emerging technologies. *Int J Multidiscip Approach Stud* 2(6):109–114
- Di Renzo M et al (2014) Spatial modulation for generalized MIMO: challenges, opportunities, and implementation. *Proc IEEE* 102(1):56–103
- Holma H, Antti T (2009) LTE for UMTS: OFDMA and SC-FDMA based radio access. Wiley, Hoboken
- ITU-R (2008) Report M. 2134, Requirements related to technical performance for IMT-Advanced radio interface(s), Approved in November 2008
- Tripathi GC et al (2015) Low cost implementation of software defined radio for improved transmit quality of 4G signals. In: IEEE communication, control and intelligent systems, Mathura
- Zheng K, Lin H, Wang W, Yang G (2005) TD-CDM-OFDM: evolution of TD-SCDMA toward 4G. *IEEE Commun Mag* 43(1):45–52

Lab-on-a-Chip

- [Communications and Networking in Droplet-Based Microfluidic Systems](#)

Large-Scale Antenna Systems (LSAS)

- [Massive MIMO](#)

Large-Scale MIMO

- [Massive MIMO](#)

Large-Scale Multiple-Antennas Channel Estimation

- [Massive MIMO Channel Estimation](#)

Lattice Reduction

- [Lattice Reduction and Its Applications in Wireless Sensors Network](#)

Lattice Reduction and Its Applications in Wireless Sensors Network

Jinming Wen

College of Information Science and Technology,
and College of Cyber Security, Jinan University,
Guangzhou, China

Synonyms

[Lattice reduction](#); [Wireless sensors networks](#)

Definitions

A *unimodular matrix* $\mathbf{U}$ is a square integer matrix whose determinant is 1 or -1 .

A *lattice* is a set of all integer linear combinations of a group of linearly independent vectors. More formally, for a given full column rank matrix $\mathbf{A} \in \mathbf{R}^{m \times n}$, the lattice $\mathcal{L}(\mathbf{A})$ generated by $\mathbf{A}$ is defined as $\mathcal{L}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} | \mathbf{x} \in \mathbf{Z}^n\}$. For example, if $\mathbf{A}$ is an $n \times n$ identity matrix, then $\mathcal{L}(\mathbf{A})$ is $\mathbf{Z}^n$. The matrix $\mathbf{A}$ is referred to as a *basis matrix* of $\mathcal{L}(\mathbf{A})$, and n is known as the *dimension* or *rank* of $\mathcal{L}(\mathbf{A})$.

Lattice reduction: for any $n \times n$ unimodular matrix $\mathbf{U}$, $\mathbf{AU}$ is also a basis matrix of $\mathcal{L}(\mathbf{A})$. Therefore, $\mathcal{L}(\mathbf{A})$ has infinitely many basis matrices if $n \geq 2$. Lattice reduction refers to the process of selecting an $n \times n$ unimodular matrix $\mathbf{U}$

such that $\mathbf{AU}$ has some “nice properties”. Usually, “nice properties” of $\mathbf{AU}$ refers to its column vectors are short and close to be orthogonal.

Wireless sensor network is a large number of geographically separately distributed sensors aiming to monitor the physical environment with a central unit that collects all the data from sensors. Wireless sensor networks are widely used to simultaneously monitor many environmental conditions, such as incident, forest fire detection, air pollution monitoring, machine health monitoring, wine production temperature, sound, and so on (Deng et al. 2017; Chen et al. 2017).

Historical Background

Lattice reduction has been studied for more than two centuries. More specifically, the idea of reduction dated back to C. F. Gauss who proposed a reduction, which is called as Gauss reduction later, in 1801 (Gauss 1801). Gauss reduction, which was defined for lattice with dimension 2, was generated to arbitrary dimension by Hermite in (Hermite 1850), and this reduction is referred to as Hermite reduction. A number of other reduction techniques were developed later; this includes Hermite-Korkine-Zolotareff (HKZ) reduction (Korkine and Zolotareff 1873), Minkowski reduction (Minkowski 1896), Brun’s reduction (Brun 1919, 1920), Lenstra-Lenstra-Lovász (LLL) reduction (Lenstra et al. 1982), and Seysen’ reduction (Seysen 1993). Among these reduction strategies, Minkowski reduction, HKZ reduction, and LLL reduction are the most popular three lattice reduction techniques, so their criteria and progresses are introduced as follows.

Minkowski reduction (Nguyen and Stehlé 2009): a basis matrix $\mathbf{A}$ is called Minkowski-reduced if for any $1 \leq i \leq n$ and $\mathbf{x} \in \mathbf{Z}^n$ with the greatest common divisor of the entries of $\mathbf{x}$ being 1, we have $\|\mathbf{Ax}\|_2 \geq \|\mathbf{A_i}\|_2$.

The first Minkowski reduction algorithm was proposed by Lagrange, but it only works for $n = 2$ (Lagrange 1773). The first Minkowski reduction algorithm for $n = 3$ and $n = 4$ were, respectively, developed in (Semaev 2001) and

(Nguyen and Stehlé 2009). Minkowski reduction algorithms for arbitrary n were independently developed in (Helfrich 1985) and (Afflerbach and Grothe 1985).

The criteria of HKZ reduction and LLL reduction techniques can be characterized by the R-factor of the QR factorization of $\mathbf{A}$. Let $\mathbf{A}$ have the QR factorization $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbf{R}^{m \times n}$ is a matrix with orthonormal columns and $\mathbf{R} \in \mathbf{R}^{n \times n}$ is an upper triangular matrix. Then HKZ reduction is formally defined as follows:

HKZ reduction: a basis matrix $\mathbf{A}$ is called HKZ reduced if the R-factor $\mathbf{R}$ of the QR factorization of $\mathbf{A}$ satisfies $|r_{ij}| \leq \frac{1}{2} |r_{ii}|$ for $1 \leq i < j \leq n$ and $|r_{ii}|$ is the length of a shortest vector of lattice $\mathcal{L}(\mathbf{R}_{i:n,i:n})$, where $\mathbf{R}_{i:n,i:n}$ denotes the submatrix of $\mathbf{R}$ with columns and rows from i to n .

HKZ reduction is often simply called as KZ reduction (see, e.g., Agrell et al. 2002), and it was also called as Hermite reduction in (Helfrich 1985). The first HKZ reduction algorithm was developed by Kannan in 1983 (Kannan 1983). Since then, a number of improved HKZ reduction algorithms have been developed, see, e.g., (Agrell et al. 2002) and (Wen and Chang 2015).

LLL reduction: a basis matrix $\mathbf{A}$ is called LLL reduced (Lenstra et al. 1982) if the R-factor $\mathbf{R}$ of the QR factorization of $\mathbf{A}$ satisfies $|r_{ij}| \leq \frac{1}{2} |r_{ii}|$ for $1 \leq i < j \leq n$ and $\delta r_{ii}^2 \leq r_{i,i+1}^2 + r_{i+1,i+1}^2$, where $\delta \in (1/4, 1]$. Note that these two conditions are, respectively, referred to as size reduction and Lovász conditions.

The LLL reduction algorithm can be found in (Lenstra et al. 1982), and it has many improvements and variants, see, e.g., (Wen and Chang 2017). The complexity of reducing an n -dimensional integer matrix is a polynomial of n (Lenstra et al. 1982). However, HKZ reduction algorithms are supposed to be NP-hard since they need to solve shortest vector problems which are suspected as NP-hard problems.

If a basis matrix $\mathbf{A}$ is HKZ reduced, then it is also LLL reduced, but the inverse usually does not hold. HKZ reduction is frequently better than the LLL reduction in shortening the lengths of basis vectors and reducing the orthogonality of basis matrices.

Key Applications

Lattice reduction has important applications in many areas, such as number theory, cryptography, Global Positioning System (GPS), and wireless sensor network. For more details, see, e.g., (Nguyen and Vallée 2010) and (Wübben et al. 2011). In these applications, lattice reduction is often used to optimally or suboptimally solving a closest vector problem (CVP) or a shortest vector problem (SVP). For a given n -dimensional vector $\mathbf{y}$ and a given full column rank matrix $\mathbf{A}$, the CVP is the problem of finding the vector in lattice $\mathcal{L}(\mathbf{A})$ that is closest to $\mathbf{y}$. Thus, mathematically, it can be expressed as $\min_{\mathbf{x} \in \mathbb{Z}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$. For a given full column rank matrix $\mathbf{A}$, the SVP is the problem of finding the shortest nonzero vector in lattice $\mathcal{L}(\mathbf{A})$. Thus, mathematically it can be expressed as $\min_{\mathbf{x} \in \mathbb{Z}^n \setminus \mathbf{0}} \|\mathbf{A}\mathbf{x}\|_2$.

Some applications of lattice reduction are briefly introduced in the following.

Applications of lattice reduction in number theory Lattice reduction has many applications in number theory. One of the applications, which is also mentioned in (Lenstra et al. 1982), is to simultaneous diophantine approximation. Specifically, for any given integer n , real numbers $y_1, y_2, \dots, y_n$ and ε which satisfies $0 < \varepsilon < 1$, the LLL reduction can find integers $x_1, x_2, \dots, x_n, \alpha$ such that $|x_i - \alpha y_i| \leq \varepsilon$ for $1 \leq i \leq n$ and $1 \leq \alpha \leq 2^{n(n+1)/4} \varepsilon^{-n}$. The integers $x_1, x_2, \dots, x_n, \alpha$ can be obtained through using the LLL reduction to reduce the basis matrix $\begin{pmatrix} \mathbf{I} & \mathbf{y} \\ \mathbf{0}^T & 2^{-n(n+1)/4} \varepsilon^{-(n+1)} \end{pmatrix}$, where $\mathbf{I}$ is the $n \times n$ identity matrix; $\mathbf{y}$ is an n -dimensional column vector whose entries are $y_1, y_2, \dots, y_n$; and $\mathbf{0}$ is an n -dimensional zero column vector. For more details, see (Lenstra et al. 1982). In addition to the above application, LLL reduction has many other applications in number theory; for more details, see, e.g., (Nguyen and Vallée 2010, Chapter 7).

Applications of lattice reduction in cryptography Lattice reduction, especially the LLL reduc-

tion, has many applications in cryptography, such as solving the problems of RSA, factoring integers, finding discrete logarithms, and solving the subset sum problem. For more details, see, e.g., (Nguyen and Vallée 2010, Chapters 10 and 11). Usually in these applications, an SVP or a CVP needs to be solved. LLL and/or HKZ reduction can be used to improve the efficiency of exactly solving SVP and CVP. It can also be used to suboptimally solve the SVP. Specifically, the first vector of the LLL-reduced matrix of $\mathbf{A}$ can be used as a suboptimal solution of the SVP. After performing the LLL reduction to $\mathbf{A}$, the Babai's nearest plane algorithm can be used to find a good suboptimal solution of the CVP.

Applications of lattice reduction in GPS In GPS, one needs to detect an n -dimensional integer vector $\mathbf{z}$ and a three-dimensional real vector $\mathbf{x}$ from the linear model $\mathbf{y} = \mathbf{A}\mathbf{z} + \mathbf{B}\mathbf{x} + \mathbf{v}$, where $\mathbf{y} \in \mathbb{R}^m$ is the difference of the observed vector and computed double-difference carrier-phase measurements; $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times 3}$ are the design matrices for ambiguity terms and baseline components, respectively; $\mathbf{z}$ and $\mathbf{x}$ are, respectively, the vectors of unknown double-difference ambiguities and unknown baseline components; and $\mathbf{v}$ is the vector of measurement noise. For more details, see (Teunissen et al. 1997) and (Hassibi and Boyd 1998). To get a maximal likelihood estimator of $\mathbf{z}$, a CVP needs to be solved. Thus the LLL reduction can be used to improve the efficiency of getting an optimal solution of the CVP and improve the effectiveness of the Babai's nearest plane algorithm, which is an effective suboptimal algorithm for the CVP.

Applications of lattice reduction in wireless sensor network Suppose that a wireless sensor network is composed of a collection of low-end data collection nodes which are connected to high-end data collection nodes through a wireless link, and there are m transmitting data collection nodes and n receiver antennas at the data gathering node. Then the received discrete time signal over a narrow-band, flat fading wireless link can be expressed as $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$, where $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{v}$ are, respectively, m -dimensional received signal,

n -dimensional transmitted signal whose entries are some discrete numbers, and n -dimensional noise signal, and $\mathbf{A} \in \mathbf{R}^{m \times n}$ is the matrix of complex fading coefficients. For more details, see (Jayaweera 2007). To get a maximal likelihood estimator of $\mathbf{x}$, a CVP needs to be solved. Thus, lattice reduction can be used to improve the efficiency or effectiveness of detecting $\mathbf{x}$.

In addition to the above applications, lattice reduction has many other applications, which includes integer optimization (see, e.g., Nguyen and Vallée 2010, Chapter 9), transportation, and so on.

Cross-References

- [Application of Machine Learning in Wireless Sensor Network](#)
- [Machine Learning in Wireless Sensor Networks for the Internet of Things](#)
- [Wireless Sensor Network Security](#)

References

- Afflerbach L, Grothe H (1985) Calculation of Minkowski-reduced lattice bases. *Computing* 35:269–276
- Agrell E, Eriksson T, Vardy A, Zeger K (2002) Closest point search in lattices. *IEEE Trans Inf Theory* 48:2201–2214
- Brun V (1919) En generalisation av kjedebrøken I. *Skr Vidensk Selsk Kristiana. Mat Nat Klasse* 6:1–29
- Brun V (1920) En generalisation av kjedebrøken II. *Skr Vidensk Selsk Kristiana. Mat Nat Klasse* 6:1–24
- Chen J, Hu K, Wang Q, Sun Y, Shi Z, He S (2017) Narrowband internet of things: implementations and applications. *IEEE Internet Things J* 4(6):2309–2314
- Deng R, He S, Cheng P, Sun Y (2017) Towards balanced energy charging and transmission collision in wireless rechargeable sensor networks. *J Commun Netw* 19(4):341–350
- Gauss CF (1801) *Disquisitiones arithmeticae*. Springer, Berlin
- Hassibi A, Boyd S (1998) Integer parameter estimation in linear models with applications to GPS. *IEEE Trans Signal Process* 46:2938–2952
- Helfrich B (1985) Algorithms to construct Minkowski reduced and Hermite reduced lattice bases. *Theory Comput Sci* 41:125–139
- Hermite C (1850) Extraits de lettres de M. Hermite à M. Jacobi sur différents objets de la théorie des nombres, deuxième lettre. *Journal für die reine und angewandte Mathematik* 40:279–290
- Jayaweera SK (2007) V-BLAST-Based virtual MIMO for distributed wireless sensor networks. *IEEE Trans Commun* 55:1867–1872
- Kannan R (1983) Improved algorithms for integer programming and related problems. In: 15th ACM symposium on theory of computing, pp 193–206
- Korkine A, Zolotareff G (1873) Sur les formes quadratiques. *Mathematische Annalen* 6:366–389
- Lagrange JL (1773) *Recherches d'arithmétique*. Nouveaux Mémoires de l'Académie de Berlin
- Lenstra AK, Lenstra HW, Lovász L (1982) Factoring polynomials with rational coefficients. *Mathematische Annalen* 261:515–534
- Minkowski H (1896) *Geometrie der Zahlen*. Teubner-Verlag
- Nguyen PQ, Stehlé D (2009) Low-dimensional lattice basis reduction revisited. *ACM Trans Algorithms, Article* 46 5(4):1–48
- Nguyen PQ, Vallée B (2010) *The LLL algorithm*. Springer, Berlin/Heidelberg
- Semaev I (2001) A 3-dimensional lattice reduction algorithm. In: *Cryptography and lattices conference*, pp 181–193
- Seysen M (1993) Simultaneous reduction of a lattice basis and its reciprocal basis. *Combinatorica* 13:363–376
- Teunissen PG, Jonge PD, Tiberius C (1997) The least-squares ambiguity decorrelation adjustment: its performance on short GPS baselines and short observation spans. *J Geod* 71:589–602
- Wen J, Chang XW (2015) A modified KZ reduction algorithm. In: 2015 IEEE international symposium on information theory (ISIT), pp 451–455
- Wen J, Chang XW (2017) GfcLLL: a Greedy selection based approach for fixed-complexity LLL reduction. *IEEE Commun Lett* 21:1965–1968
- Wübben D, Seethaler D, Jaldén J, Matz G (2011) Lattice reduction: a survey with applications in wireless communications. *IEEE Signal Process Mag* 28:70–91

Learning-Based Secure Mobile Offloading

Liang Xiao

Department of Communication Engineering,
Xiamen University, Xiamen, China

Definition

Learning-based secure offloading strategy applies the reinforcement learning algorithms to derive the optimal data offloading and defense policy for the mobile devices to counter the potential smart

attackers in a mobile computing system, without any preliminary knowledge of the environmental parameters and the attack model.

Historical Background

With the proliferation of cloud-based mobile services, mobile devices such as smartphones and tablets can offload their applications and data to the cloud to improve user experience in terms of longer battery lifetime, larger data storage, faster processing speed, and more powerful security services (Xiao et al. 2017). However, data offloading to the cloud via access points (APs) or base stations (BSs) is vulnerable to various types of attacks, such as spoofing, eavesdropping, and jamming (Kumar and Lu 2010). A smart attacker can use smart and programmable radio devices such as Universal Software Radio Peripherals (USPRs) or the Wireless Open-Access Research Platform (WARP) developed by Rice University (Murphy et al. 2006) to throw multiple types of attacks. Compared with traditional single mode attacks, mobile offloading is more vulnerable to smart attacks, as a smart attacker flexibly chooses its attack mode and strength according to the ongoing offloading transmissions and radio environment. For example, a smart attacker can throw jamming attacks if it is close to the serving AP and can efficiently block the offloading or can send spoofing signals with the MAC address of the mobile device.

Although the detection of spoofing and jamming attacks has been well studied (Mukherjee and Swindlehurst 2010; Hu et al. 2015), mobile devices still suffer from time and energy loss due to false alarms and security loss resulting from missed detection of attacks. Security agents at APs or BSs can apply both physical-layer and higher-layer security mechanisms (He et al. 2012) to protect the offloading process. More specifically, the security agent can apply an advanced security mechanism, possibly by processing the offloading data again or changing the session keys, at the cost of higher processing and transmission overhead. The offloading rate of the mobile device is chosen according to the radio

environment, e.g., a mobile device has to process the data locally under strong jamming or heavy traffic at the AP or BS (Yang et al. 2015; Wang et al. 2013; Li et al. 2001; Chen 2015; Xiao et al. 2018c,d).

While game theory has been used to study the interactions between attackers and mobile devices (Mukherjee and Swindlehurst 2010; Hu et al. 2015; Xiao 2015; Xiao et al. 2018a,b,e), mobile offloading usually involves three players: a mobile device, a smart attacker, and a security agent. The secure mobile offloading game is investigated in Xiao et al. (2016), in which a smart attacker can perform multiple types of attacks, including spoofing and jamming, a mobile device determines its offloading rate, and a security agent at the AP protects the offloading with two modes: the fast mode that applies the physical-layer security and the safe mode that applies both the physical-layer and higher-layer security mechanisms. In the dynamic offloading games, a mobile device has difficulty obtaining all the system parameters, such as the channel condition and transmit powers of the attacker in dynamic environments needed to derive the optimal offloading strategy. As an important reinforcement learning technique, Q-learning can be applied by the mobile device to derive the optimal offloading strategy via trials.

Foundations

The secure offloading system consists of a mobile device ($\mathcal{S}$) under the protection of a security agent ($\mathcal{D}$) at the serving access point (AP) and a potential smart attacker ($\mathcal{E}$). The mobile device offloads its applications and data to the cloud via the AP. The offloading rate of the mobile device, denoted by $x \in [0, 1]$, is defined as the proportion of the data that is sent to the serving AP. If $x = 0$, the mobile device processes all the data locally, while $x = 1$ means that all the data are processed in the cloud. The offloading rate is chosen as a tradeoff among the fast processing in the cloud, the radio transmission cost, and the risks of being attacked during offloading.

The smart attacker chooses its attack mode against the mobile device, denoted by $y \in \{0, 1, 2\}$, which corresponds to no attack, spoofing, and jamming, respectively. More specifically, the attacker sends spoofing signals with the mobile device's identity when $y = 1$, transmits jamming signals with jamming power P_E to block the offloading when $y = 2$, and keeps silent when $y = 0$.

The security agent operates in two modes: the safe mode which combines a physical-layer security method such as the channel-based spoofing detector (Xiao et al. 2009) and advanced higher-layer security mechanisms such as the authentication technique presented in He et al. (2012) for attack detection at the AP when $z = 1$ and a fast mode in which only the physical-layer security mechanism is used if $z = 0$. The extra cost of the safe mode over the fast mode denoted by β is positive by definition, and the cost of the fast mode is ignored. Let C_y^z denote the attack cost of the mobile device if the attacker adopts attack mode y and the security agent protects the AP via mode z .

The channel gain between the AP and the mobile device is modeled as a Markov chain. More specifically, the channel gain during the offloading at time n , denoted by $h_T^n \in \mathbf{H}_T = \{H_T^l\}_{1 \leq l \leq N_T}$, is quantized into N_T levels, where H_T^l is the l -th level of the channel gain with $H_T^m < H_T^k, \forall 1 \leq m < k \leq N_T$. The transition probability of the channel gain h_T^n is defined as $p_{m,k} = \Pr(h_T^{n+1} = H_T^k | h_T^n = H_T^m)$. Similarly, the channel gain between the attacker and the AP, denoted by $h_E^n \in \mathbf{H}_E = \{H_E^j\}_{1 \leq j \leq N_E}$, is modeled as a Markov chain with N_E states with transition probabilities denoted by $q_{m,k} = \Pr(h_E^{n+1} = H_E^k | h_E^n = H_E^m)$. For simplicity of notation, the time index of the channel gain is omitted if no confusion results.

It is assumed that both the mobile device and the security agent make decisions to maximize the same utility denoted by u , which depends on the attack mode, the channel gains, and the transmit power of the offloading signals denoted by P_T . Thus, the utility of the mobile device is given by:

$$u(x, y, z) = \begin{cases} x (\log(1 + P_T h_T) - \sigma P_T) - z\beta, & y = 0 \\ x (\log(1 + P_T h_T) - \sigma P_T) - C_1^z - z\beta, & y = 1 \\ x \left(\log \left(1 + \frac{P_T h_T}{1 + P_E h_E} \right) - \sigma P_T \right) - C_2^z - z\beta, & y = 2 \end{cases} \quad (1)$$

where $\sigma > 0$ is the positive unit transmission cost.

In this game, the utility of the attacker denoted by $u_{\mathcal{E}}$ is given by:

$$u_{\mathcal{E}}(x, y, z) = -u(x, y, z). \quad (2)$$

A dynamic mobile offloading game is formulated to model the repeated offloading interactions in dynamic radio environments, which consists of a mobile device, smart attacker, and security agent that are unaware of the system parameters such as the attack cost, the offloading gain, and the current channel condition. Based on Q-learning, the mobile device determines its offloading rate based on the actions of its opponents and the channel condition in the last time slot. To this end, the mobile device observes

the system state denoted by $\mathbf{s}$ at time n given by $\mathbf{s}^n = [y^{n-1}, z^{n-1}, h_T^{n-1}]$. In this game, the offloading rate of the mobile device is quantified into L levels, i.e., $x^n \in \{i/(L-1)\}_{0 \leq i \leq L-1}$ for simplicity.

Let $Q(\mathbf{s}, x)$ denote the quality function of the mobile device for offloading rate x and state $\mathbf{s}$. The value function denoted by $V(\mathbf{s})$ represents the maximum value of the quality function at state $\mathbf{s}$. The mobile device updates its Q-function based on its utility u and the value function as follows:

$$Q(\mathbf{s}^n, x^n) \leftarrow (1 - \alpha) Q(\mathbf{s}^n, x^n) + \alpha \left(u(\mathbf{s}^n, x^n) + \delta V(\mathbf{s}^{n+1}) \right) \quad (3)$$

$$V(\mathbf{s}^n) = \max_x Q(\mathbf{s}^n, x), \quad (4)$$

where $\alpha \in (0, 1]$ is a learning factor indicating the weight of the current estimate of the function Q in the update of the quality function, and the discount factor δ represents the uncertainty of the mobile device about the future rewards.

By applying the ε -greedy policy (Sutton and Barto 1998), the mobile device chooses its offloading rate x^n to maximize its current Q -function with a high probability $1 - \varepsilon$, while the other $L - 1$ rates are taken with equal probability, i.e.,

$$\Pr(x^n) = \begin{cases} 1 - \varepsilon, & x^n = \arg \max_x Q(s^n, x) \\ \frac{\varepsilon}{L-1}, & o.w. \end{cases} \quad (5)$$

Key Applications

Learning-based secure mobile offloading strategy provides the mobile device with intelligent defense mechanism against smart attacks in the mobile computing system, which balances the security performance and the energy consumption of the devices by adjusting the offloading data rate and the defense mode. Even without knowing the environmental parameters, the mobile devices can apply it to derive the optimal policy to improve the secure data communication efficiency via trial and errors.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant 61671396.

References

- Chen X (2015) Decentralized computation offloading game for mobile cloud computing. *IEEE Trans Parallel Distrib Syst* 26(4):974–983
- He D, Chen C, Chan S, Bu J (2012) Secure and efficient handover authentication based on bilinear pairing functions. *IEEE Trans Wirel Commun* 11(1):48–53
- Hu P, Li H, Fu H, Cansever D, Mohapatra P (2015) Dynamic defense strategy against advanced persistent threat with insiders. In: *Proceedings of IEEE international conference on computer communications (INFOCOM)*, pp 747–755
- Kumar K, Lu YH (2010) Cloud computing for mobile users: can offloading computation save energy? *IEEE Comput* 43(4):51–56
- Li Z, Wang C, Xu R (2001) Computation offloading to save energy on handheld devices: a partition scheme. In: *Proceedings of ACM international conference on compilers, architecture, and synthesis for embedded systems*, pp 238–246
- Mukherjee A, Swindlehurst AL (2010) Optimal strategies for countering dual-threat jamming/eavesdropping-capable adversaries in MIMO channels. In: *Proceedings of IEEE military communications conference*, pp 1695–1700
- Murphy P, Sabharwal A, Aazhang B (2006) Design of WARP: a wireless open-access research platform. In: *IEEE signal processing conference, 2006 14th European*, pp 1–5
- Sutton R, Barto AG (1998) *Reinforcement learning: an introduction*. MIT Press, Cambridge
- Wang Y, Lin X, Pedram M (2013) A nested two stage game-based optimization framework in mobile cloud computing system. In: *Proceedings of IEEE international symposium on service oriented system engineering*, pp 494–502
- Xiao L (2015) *Anti-jamming transmissions in cognitive radio networks*. Springer. ISBN:978-3-319-24290-3
- Xiao L, Greenstein LJ, Mandayam NB, Trappe W (2009) Channel-based spoofing detection in frequency-selective rayleigh channels. *IEEE Trans Wirel Commun* 8(12):5948–5956
- Xiao L, Xie C, Chen T, Dai H, Poor HV (2016) A mobile offloading game against smart attacks. *IEEE Access* 4:2281–2291
- Xiao L, Li Y, Huang X, Du X (2017) Cloud-based malware detection game for mobile devices with offloading. *IEEE Trans Mobile Comput* 16(10):2742–2750
- Xiao L, Chen T, Xie C, Dai H, Poor HV (2018a) Mobile crowdsensing games in vehicular networks. *IEEE Trans Veh Technol* 62(2):1535–1545
- Xiao L, Li Y, Han G, Dai H, Poor HV (2018b) A secure mobile crowdsensing game with deep reinforcement learning. *IEEE Trans Inf Forensics Secur* 13(1):35–47
- Xiao L, Wan X, Dai C, Du X, Chen X, Guizani M (2018c) Security in mobile edge caching with reinforcement learning. *IEEE Wirel Comm* 25(3):116–122 Mag
- Xiao L, Wan X, Lu X, Zhang Y, Wu D (2018d) IoT security techniques based on machine learning. *IEEE Sig Process Mag* 35(5):41–49
- Xiao L, Xu D, Mandayam N, Poor HV (2018e) Attacker-centric view of a detection game against advanced persistent threats. *IEEE Trans Mobile Comput* 17(11):2512–2523
- Yang Z, Niyato D, Wang P (2015) Offloading in mobile cloudlet systems with intermittent connectivity. *IEEE Trans Mobile Comput* 14(12):2516–2529

Lens Antenna Array

Tian Xie¹, Linglong Dai², Yongming Huang³, and Ming Xiao⁴

¹Department of Electronic Engineering, Tsinghua University, Beijing, China

²Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China

³School of Information Science and Engineering, Southeast University, Nanjing, China

⁴Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Beamspace MIMO

Definitions

Lens antenna array refers to a special category of antenna arrays in millimeter wave multiple-input multiple-output (MIMO) systems. The most distinguishing characteristic of lens antenna arrays is that there is an electromagnetic lens placed between the antenna array and the coming direction of wireless signals. Intuitively, the lens antenna arrays focus the signals from a certain spatial direction onto a particular antenna element. From the perspective of signal processing, the function of the electromagnetic lens can be interpreted as a phase shifting network, which transforms the spatial MIMO channel into the beamspace channel.

Historical Background

Lens antenna array is proposed very recently, mainly coming with the research on millimeter wave MIMO systems. The earliest representative work on lens antenna array that could be found in literature was conducted by A. M. Sayeed and J. Brady from University of Wisconsin, Madison, in

2013 (Brady et al. 2013). In (Brady et al. 2013), the authors built a general computational framework for MIMO systems with lens antenna arrays (namely, beamspace MIMO systems), based on which the link capacity was analyzed. To verify the accuracy of their models, authors in (Brady et al. 2013) also compared results from theoretical analysis and prototype measurements, where a fairly high consistency was observed. Following their work in (Brady et al. 2013), A. M. Sayeed's group has continued to generalize the initial work in single-user line-of-sight scenario to multipath (Song et al. 2013) and multiuser scenario (Brady and Sayeed 2014).

Shortly after (Brady et al. 2013), another group from National University of Singapore also published their work on lens antenna array (Zeng et al. 2014, 2017; Zeng and Zhang 2016), where more emphasis have been put on the transmission scheme design in MIMO systems with lens antenna arrays. In (Zeng et al. 2014), Y. Zeng and R. Zhang applied an electromagnetic lens enabled beamspace MIMO system in a single-cell multiuser uplink scenario. A new path division multiplexing (PDM) paradigm has been proposed in (Zeng and Zhang 2016) based on the beamspace MIMO, where different data streams are transmitted over different paths. Furthermore, Y. Zeng and R. Zhang continued to generalize the single-user PDM paradigm to the multiuser case under the framework of path division multiple access (PDMA) in (Zeng et al. 2017).

More recently, various studies on MIMO systems with lens antenna arrays have sprang up. Though MIMO systems with lens antenna arrays can achieve obvious performance gains, such gains are highly dependent on accurate channel state information (CSI) at transceivers. To address the CSI acquisitions issue, X. Gao et al. in 2017 proposed a compressive sensing-based channel estimation approach with low pilot overhead and high accuracy. To improve the spectral efficiency of the system, B. Wang et al. in 2017 investigated the combination of MIMO systems with lens antenna array and non-orthogonal multiple access (NOMA) techniques, i.e., beamspace MIMO NOMA. Compared with conventional beamspace MIMO, beamspace MIMO NOMA is able to serve multiple users

simultaneously in the same beam, which can obviously improve the achievable sum rate of the system.

Foundations

Lens antenna arrays are usually utilized in the high-frequency band MIMO systems (e.g., millimeter wave (mmWave) MIMO or Tera Hz (THz) MIMO), to reduce the signal processing dimensions. In high-frequency band communications, MIMO techniques with large-scale antenna arrays are usually utilized to compensate the severe propagation loss. However, as the number of antennas in MIMO systems increases, it is very challenging to operate key signal processing procedures such as precoding and channel estimations. The most prominent reason for this problem is that the conventional fully digital transceiver architecture requires one dedicated radio frequency (RF) chain (including high-resolution digital-to-analog converters) for each antenna. Such configuration brings prohibitively high energy consumption and hardware cost, since the RF chains will be power-hungry and costly in the mmWave and THz frequency.

Intuitively, the function of lens antenna arrays can be interpreted as focusing the signal coming from each path onto a certain antenna element. From the perspective of signal processing, the lens antenna arrays perform a uniform spatial Fourier transformation on the channel, i.e., transfer the conventional spatial MIMO channel into the beamspace channel. In the beamspace channel, each channel element corresponds to the gain of one path toward certain direction. Thanks to the limited scattering characteristics in mmWave or THz signal propagations, the beamspace channel presents an obvious sparse structure. Basically, only those channel elements with obvious power need to be selected. In other words, the communication in beamspace channel only occurs in a subspace, which indicates that the dimension of MIMO systems can be much reduced.

The dimension reducing ability of lens antenna arrays mainly depends the sparsity level

of the beamspace channels, i.e., the number of available paths in signal propagations. This explains why lens antenna arrays are more suitable for high-frequency communications, where the number of paths is usually small due to the high path loss. In lower frequencies (e.g., sub-6 GHz) where the channel demonstrates a denser scattering, using lens antenna array will hardly reduce the system dimensions.

Key Applications

Lens antenna arrays are usually utilized in the high-frequency band MIMO systems to reduce the system dimensions. With a lower dimension, the enormous energy consumption and hardware cost due to huge RF networks or phase shifter networks can be greatly saved.

Cross-References

- [Hybrid Precoding](#)
- [Millimeter-Wave \(mmWave\) Multiple-Input Multiple-Output \(MIMO\) Technique](#)
- [Millimeter Wave NOMA](#)

References

- Brady J, Sayeed AM (2014) Beamspace MU-MIMO for high-density gigabit small cell access at millimeter-wave frequencies. In: IEEE signal processing and advanced wireless communications (SPAWC). IEEE, pp 80–84
- Brady J, Nader B, Sayeed AM (2013) Beamspace MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements. IEEE Trans Antennas Propag 61(7):3814–3827
- Gao X, Dai L, Han S, I CL, Wang X (2017) Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array. IEEE Trans Wirel Commun 16(9):6010–6021
- Song GH, Brady J, Sayeed AM (2013) Beamspace MIMO transceivers for low-complexity and near-optimal communication at mm-wave frequencies. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4394–4398
- Wang B, Dai L, Wang Z, Ge N, Zhou S (2017) Spectrum and energy-efficient beamspace MIMO-NOMA for

- millimeter-wave communications using lens antenna array. *IEEE J Sel Areas Commun* 35(10):2370–2382
- Zeng Y, Zhang R (2016) Millimeter wave MIMO with lens antenna array: a new path division multiplexing paradigm. *IEEE Trans Commun* 64(4):1557–1571
- Zeng Y, Zhang R, Chen Z (2014) Electromagnetic lens-focusing antenna enabled massive MIMO: performance improvement and cost reduction. *IEEE J Sel Areas Commun* 32(6):1194–1206
- Zeng Y, Yang L, Zhang R (2017) Multi-user millimeter wave MIMO with single-sided full-dimensional lens antenna array. In: *IEEE international conference on communications (ICC)*. IEEE, pp 1–6

License-Assisted Access. (Not Licensed-...)

- [Cellular Unlicensed Spectrum Technology](#)

Licensed-Assisted Access (LAA)

- [Effective Utilization of Unlicensed Spectrum](#)

Light-Emitting Diode

- [Indoor Visible Light Communication Networks for Camera-Based Mobile Sensing](#)

Linear Modulation

- [Equalization Techniques for Single-Carrier Modulations](#)

Link Assignment

- [Contact Plan Design for Space Disruption/Delay-Tolerant Networks](#)

Live Video Streaming in Wireless P2P Networks

Bo Zhang, Yangjie Cao, and Ying Han
School of Software and Applications
Technologies, Zhengzhou University,
Zhengzhou, China

Synonyms

[Wireless Ad Hoc Networks](#)

Definitions

Primary network refers to a wide-area wireless access network provided by mobile communication companies or other ISPs, such as 3G, 4G, etc.

Secondary network refers to a short-range device-to-device connections such as Wi-Fi Direct, Bluetooth, NFC, etc.

Historical Background

In recent years, the capacity of mobile devices, including both networking and computational, has been kept improving. However such development still falls short when facing increasing requirement from various kind of new services developed.

To address such inconsistency, the term “cloud” has been proposed and swiftly popularized in both academia and industry. Cloud off-loads a great deal of computational tasks from mobile devices to a group of powerful servers remotely located. Cloud enables low-capacity devices to perform computational-intensive tasks, by trading-off precious bandwidth resources and potentially long delay.

It is often the case that a mobile device’s task is only valid and useful locally, i.e., within a small proximity of the device. Therefore, those tasks are best performed at the device end. With the development of mobile peer-to-peer connectivity, Mobile devices are now able to connect with each

other to form a high-speed wireless P2P network at the edge to perform computational-intensive tasks by sharing their resources.

Among those tasks, one promising type is video live streaming. With the development of video formats, a single device can rarely stream a high-quality (e.g., high definition, high frame rate, etc.) video with stable and satisfactory quality. For regionally popular videos, it is possible that a group of nearby mobile devices interested in the content form a P2P network and collaboratively complete the streaming and sharing task.

Introduction

With the advances in multimedia capability of wireless devices, live video streaming to handheld devices has become a reality. Conventional streaming approaches utilize the client-server model, in which each mobile station (MS) independently *pulls* streams from the server

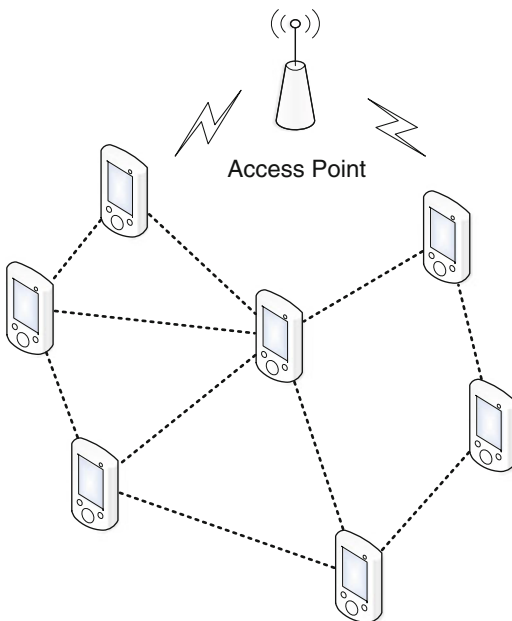
via an access point (AP) or base station (BS). However such approach suffers from several limitations including low scalability and coverage blind spots.

To overcome the above limitations, wireless cooperative streaming has been proposed. This is motivated by the much success of peer-to-peer (P2P) streaming over the Internet. In wireless P2P streaming, devices form a wireless multihop network and cooperatively exchange video data with each other to achieve efficient stream broadcasting. This architecture is becoming possible by the fact that more and more wireless devices nowadays are equipped with not only multiple wireless interfaces (e.g., 3G, Wi-Fi, and Bluetooth) as well as antennas but also large memory and high processing capabilities.

Figure 1 shows the network considered for mobile P2P live streaming, where dashed lines indicate the underlay connectivities. There is at least one *source node* in the network who either generates video content or gets it from the streaming server via a long-range *primary* channel such as a 3G or 4G network. The sources then share the data with their neighbor via a *secondary* channel. The receiving nodes then further share their content to their neighbors.

Such a cooperative wireless live video streaming network needs to be efficient in that given limited resources, it needs to achieve good service quality. More specifically, the network needs to be efficient in the following aspects:

- *Optimized overlay* The streaming overlay needs to be efficient in terms of low control overhead and low traffic redundancy given sufficient user coverage, irrespective of the network dynamics.
- *High playback quality* Video distortion at the client side is an important QoS metric for video streaming. Given limited resources, the streaming network should deliver video stream to users with high playback quality.
- *Efficient loss recovery* Packet loss is inevitable in wireless video streaming due to dynamic channel condition. These losses must be recovered with low recovery traffic.



Live Video Streaming in Wireless P2P Networks, Fig. 1 Mobile peer-to-peer network in consideration. A subset of mobile devices requests video stream from content server via primary channel through access point/base station and then share their stream with nearby neighbors via a secondary channel

To understand how the above discernment may be achieved, the characteristics of wireless multihop network need to be first reviewed. Wireless multihop network bears the following distinctive characteristics:

- *Broadcast nature* Each wireless transmission will be overheard by neighbors within range. This broadcast nature brought several unique properties including high transmission efficiency, high loss recovery efficiency, high spatial diversity and loss recovery efficiency (Li et al. 2013), per-hop media contention, and potentially high interference/delay.
- *Network dynamicity* Wireless multihop network is dynamic due to peer churn, making schemes designed for relatively static networks not applicable.
- *Scarce network resource* Network resources in wireless environment, including mainly device computational power, network bandwidth, and device energy, are often severely limited and cannot afford the high control overhead as in wired P2P.
- *Limited “parent” candidates* In wireless environment a client can only get supplied from a few neighboring nodes.
- *Different natures of loss* Losses in wireless network are generally not caused by congestion as in wired network, but rather due to signal errors as a result of noise, interference, or transient disconnection. In this case the congestion avoidance mechanisms such as the one used in TCP will only harm the effective throughput.

In the remainder of this chapter, the aforementioned three properties of a wireless multihop streaming network will be discussed, with proposed approaches addressing them introduced.

Overlay Optimization

Appropriate overlay design crucially affects the performance of the streaming network. Wireless P2P overlays generally fall into one

of three major categories, namely, unstructured, structured, and hybrid overlays, based on whether transmissions follow predetermined routes.

The main characteristic of an **unstructured overlay** is that the routing of each data piece is determined independently on the fly at each hop. The motivation is primarily its ability of dealing with network dynamics, as well as its simplicity and distributed nature. An efficient mesh-based mobile P2P live streaming protocol called **COSMOS** has been proposed in Leung and Chan (2007). In COSMOS a small number of mobile peers pull a video description from the server via an AP, and then each determines whether to push the content to its neighbors. The heuristic applied is the fraction of puller’s one-hop neighbors that could benefit from such rebroadcast. Discovering and selecting *partner set* are proposed in **P2PMLS** (Jinfeng et al. 2007). A new joiner first checks if there is any neighbor peers or multihop nodes receiving the same data. If such neighbors are found, a subset of them is chosen to be stream suppliers. Otherwise the peer becomes a puller itself. The works in Wu and He (2011) introduces *preference score* of a neighbor node, which measures the mobility pattern similarity as well as the load of the neighbor node. *Tendency score* is also proposed to quantify the “tendency” that a neighbor may become a good supplier. Li et al. proposed a distributed scalable video (SVC) multicast algorithm that jointly optimizes video quality and network traffic (Li et al. 2012). The scheme finds multiple disjoint paths and a backup path for each destination. Another similar optimized flow selection algorithm is proposed in Abdel Khalek and Dawy (2012).

In **structured overlay**, a broadcast spanning tree is continuously maintained so that the routes to all the clients via this tree are readily available for streaming. The major advantage of such proactive, tree-based overlay construction is its shorter routing delay and less redundant transmissions. However, a higher control overhead is normally required and the overlay continuously updated due to network dynamics, even when there is no data.

In single-tree overlay construction, mobile nodes form a broadcast tree spanning all the nodes. Much research work on mobile P2P streaming are single-tree based. Chen and Kao in (2013) employs a game-based parent selection algorithm called **GB-BTC** for tree construction. Each node in GB-BTC will prefer to select the neighbor node with maximum number of children as well as low expected transmission counts (ETXs) as its parent. Recently, Chang et al. in (2014) proposed **PNCB** to build a broadcast tree for each video broadcasting source. The work suggests that the number of transmissions required at any node is lower bounded by the maximum number of trees that select a common outgoing link from this node. Consequently, a neighboring node is preferred to be the parent if doing so does not increase the neighbor's max out-degree.

In single-tree approach, leaf nodes often suffer from high loss due to error propagation and the delay between unexpected disconnections and subsequent parent renegotiations. Multi-tree approach tries to address this by constructing multiple trees for concurrent use. It usually makes use of multiple description coding (MDC) or scalable video coding (SVC) and constructs multiple trees each for a description/layer of the video.

Hybrid overlay topology, often referred to as structured mesh, is to combine the strengths of both unstructured mesh approach and tree approach. Two types of hybrid architectures are introduced, namely, interleaved hybrid and tiered hybrid (Kubo et al. 2010). Interleaved hybrid topology tries to add additional links over a tree structure. Tiered hybrid topology constructs a tree with high relay-ability nodes, which supports the mesh network formed by the rest of the nodes.

Another hybrid topology, called **Local-Tree** (Zhang et al. 2011), aims at optimizing an unstructured *mesh* with regional tree structures. In LocalTree, new joiners initially form an unstructured mesh. Relatively stable node groups (based on link transmission success rate and node relay probability) are identified distributedly. Within each node group, a locally optimized tree is computed.

Packet Loss Recovery

In wireless video broadcasting, packet loss often occurs due to the dynamic nature of wireless channel (i.e., random bit errors, burst errors, or transient outages). Timely loss recovery is hence important to achieve good service quality. In the traditional approach, a mobile station (MS) with packet loss, the so-called *errored MS*, requests the server for retransmission via a separate backward channel. There has been much work on server retransmission-based error recovery for wireless video by feeding back loss information to the server (Chachulski et al. 2007; Schier and Welzl 2012). A common assumption for all the above work is a feedback channel to the server, so that ACKs or NAKs can be reported back to trigger retransmission at the server. This is not scalable to a large pool of MSs due to the bandwidth limitation of the wireless channel and at the server.

With the development of network coding (NC), another approach lets the server perform network coding on source packets and broadcast coded packets to clients. This approach requires the knowledge of network loss condition and is often designed for certain worst-case loss. There have been many erasure-resilient coding algorithms studied, such as Reed-Solomon code and fountain codes (e.g., Tornado code, Raptor code, or Luby transform (LT) code). Much work has been done using FEC to recover errors (Lee et al. 2011; Nguyen et al. 2011). FEC is normally designed for a certain worst-case loss rate, which uses bandwidth unnecessarily in a low-loss condition and is insufficient to recover loss in a high-loss condition.

Given that nowadays MSs are often equipped with a (free) secondary broadcast channel (e.g., Wi-Fi or Bluetooth), such channel may be used for cost-effective cooperative error recovery. Cooperative loss recovery has been studied in Alnuweiri et al. (2012); Liu et al. (2015). Source packet sharing is adapted in most of these works, reducing the recovery effectiveness. In many proposed NC-based approach, cooperative recovery is *unicast* based, limiting the effectiveness of recovery packets to a single receiver.

With the above insights and inspired by Liu et al. (2011b), the authors proposed **BOPPER** in Zhang et al. (2014), which combines broadcast-based cooperative recovery with linear network coding. BOPPER is therefore adaptive to heterogeneous loss rates. It is complementary to FEC and can be applied *after* FEC to reduce transmission overhead and to achieve lower residual loss rate. In BOPPER, a node generates parity packets based on the possibly partial group of source packets received. When the number of received parity packets received from other peers is no fewer than the number of lost source packets, these source packets can be recovered with very high probability. An MS generates parity packets based on the error status of its neighbors, and hence the number of parity packets generated by an MS adapts local loss condition. An MS may generate parity packets based on the source packets it received via the primary channel. Furthermore, by employing NC for local error recovery, MSs with different losses may use the same parity packets for recovery, which effectively reduces recovery traffic in the secondary channel. Neighboring nodes inform each other by broadcasting NAK messages periodically. In order to reduce control overhead, NAK further proposes NAK suppression, so that latter NAKs can be safely suppressed if the information is contained by earlier ones.

Bandwidth Utilization

Novel video formats, such as VR, AR, 360 video, and free-viewpoint video (FVV), enable new and compelling viewing experience (Liu et al. 2011a). In an implementation of FVV, *anchor views* of a scene are captured simultaneously in texture-plus-depth format by a discrete array of cameras. A user receiving some of these anchors may either play back an anchor or render a *virtual view* (i.e., views not captured by any camera) given his/her available anchors using depth-image-based rendering (DIBR).

In this section, a group of wireless users interested in a live FVV is considered. They pull the anchor streams from a remote server (through, e.g., 4G data network). As wireless bandwidth is precious, how to efficiently utilize the bandwidth to provide the best possible received video quality to the users becomes a challenging problem. One simple approach is that, upon each view switch at a user, the server renders the requested view for the user. In Miao et al. (2011), a cloud is employed to compute and render each requested view. The work in De Raffaele and Debono (2010) studies view-switching prediction to reduce traffic and computation load at the server for each view switch.

Streaming all the anchors to users for them to render all the views they need would consume prohibitively large server and network bandwidths. Streaming only a subset of anchors to each user may lead to severely degraded video quality due to increased anchor distances. To address this issue, network coding and cooperative sharing techniques may be utilized. The work termed **PAFV** in Zhang et al. (2016) considers the scenario where clients collaboratively pull anchors and share NC packets formed by the pulled anchors. Given the distortion function and spatial popularity, PAFV optimizes anchor pulling and NC generation in each client, in order to minimize the overall expected video distortion, subject to bandwidth constraints. PAFV employs a max-clique algorithm to solve the anchor sharing problem.

Conclusion

The broadcast nature of wireless transmissions and topology dynamics of wireless cooperative networks bring challenges to the overlay design of a P2P streaming system over such a network. Different types of overlay designs, namely, unstructured, structured, and hybrid approaches are discussed. In order to repair the lost packets in wireless video broadcasting, efficient cooperative loss recovery schemes are reviewed. For new video paradigms such as 360 video, VR

video, and free-viewpoint video, efficient streaming schemes are checked and discussed.

References

- Abdel Khalek A, Dawy Z (2012) Energy-efficient cooperative video distribution with statistical QoS provisions over wireless networks. *IEEE Trans Mobile Comput* 11(7):1223–1236
- Alnuweiri H, Rebai MR, Beraldi R (2012) Network-coding based event diffusion for wireless networks using semi-broadcasting. *Ad Hoc Netw* 10(6):871–885. ISSN 1570-8705
- Chachulski S, Jennings M, Katti S, Katabi D (2007) Trading structure for randomness in wireless opportunistic routing. In: *Proceedings ACM SIGCOMM, Kyoto*, pp 169–180. ACM. ISBN 978-1-59593-713-1
- Chang C-H, Kao J-C, Chen F-W, Cheng SH (2014) Many-to-all priority-based network-coding broadcast in wireless multihop networks. In: *Wireless telecommunications symposium (WTS), Washington, DC*, pp 1–6
- Chen F-W, Kao J-C (2013) Game-based broadcast over reliable and unreliable wireless links in wireless multihop networks. *IEEE Trans Mobile Comput* 12(8):1613–1624
- De Raffaele C, Debono CJ (2010) Applying prediction techniques to reduce uplink transmission and energy requirements in mobile free-viewpoint video applications. In: *Proceedings of IEEE MMEDIA, Athens/Glyfada*, pp 55–60
- Jinfeng Z, Jianwei N, Rui H, Jianping H, Limin S (2007) P2P-leveraged mobile live streaming. In: *21st international conference on advanced information networking and applications workshops (AINAW'07), Niagara Falls, Ontario*, vol 2, pp 195–200
- Kubo H, Shinkuma R, Takahashi T (2010) Topology control using multi-dimensional context parameters for mobile P2P networks. In: *IEEE 71st vehicular technology conference (VTC 2010-Spring), Taipei*, pp 1–5
- Lee J-W, Chen C-L, Horng M-F, Kuo Y-H (2011) An efficient adaptive FEC algorithm for short-term quality control in wireless networks. In: *13th international conference on advanced communication technology (ICACT), Gangwon-Do*, pp 1124–1129
- Leung M-F, Chan S-HG (2007) Broadcast-based peer-to-peer collaborative video streaming among mobiles. *IEEE Trans Broadcast Special Issue on Mobile Multimedia Broadcast* 53(1):350–361
- Li C, Xiong H, Zou J, Chen CW (2012) Distributed robust optimization for scalable video multirate multicast over wireless networks. *IEEE Trans Circuits Syst Video Technol* 22(6):943–957
- Li Y, Zhang Z, Wang C, Zhao W, Chen H-H (2013) Blind cooperative communications for multihop ad hoc wireless networks. *IEEE Trans Veh Technol* 62(7):3110–3122. ISSN 0018-9545. <https://doi.org/10.1109/TVT.2013.2256475>
- Liu Z, Cheung G, Ji Y (2011a) Distributed source coding for WWAN multiview video multicast with cooperative peer-to-peer repair. In: *2011 IEEE international conference on communications (ICC)*, pp 1–6. <https://doi.org/10.1109/icc.2011.5962865>
- Liu Z, Cheung G, Ji Y (2011b) Distributed markov decision process in cooperative peer-to-peer repair for WWAN video broadcast. In: *2011 IEEE international conference on multimedia and expo*, pp 1–6. <https://doi.org/10.1109/ICME.2011.6012176>
- Liu Z, Cheung G, Chakareski J, Ji Y (2015) Multiple description coding & recovery of free viewpoint video for wireless multi-path streaming. *IEEE J Sel Topics Signal Process* 9(1):151–164
- Miao D, Zhu W, Luo C, Chen CW (2011) Resource allocation for cloud-based free viewpoint video rendering for mobile phones. In: *Proceedings ACM MM, Scottsdale*, pp 1237–1240
- Nguyen HDT, Tran L-N, Hong E-K (2011) On transmission efficiency for wireless broadcast using network coding and fountain codes. *IEEE Commun Lett* 15(5):569–571. ISSN 1089-7798
- Schier M, Welzl M (2012) Optimizing selective ARQ for H.264 live streaming: a novel method for predicting loss-impact in real time. *IEEE Trans Multimedia* 14(2):415–430. ISSN 1520-9210. <https://doi.org/10.1109/TMM.2011.2178235>
- Wu S, He C (2011) QoS-aware dynamic adaptation for cooperative media streaming in mobile environments. *IEEE Trans Parallel Distrib Syst* 22:439–450
- Zhang B, Chan S-HG, Cheung G, Chang E (2011) Local-Tree: an efficient algorithm for mobile peer-to-peer live streaming. In: *Proceedings of IEEE international conference on communications (ICC), Kyoto*, pp 1–5
- Zhang B, Chan S-HG, Cheung G (2014) Peer-to-peer error recovery for wireless video broadcasting. In: *Peer-to-peer networking and applications*. Springer, pp 1–13. <https://doi.org/10.1007/s12083-014-0297-8>
- Zhang B, Liu Z, Chan S-HG, Cheung G (2016) Collaborative wireless freeview video streaming with network coding. *IEEE Trans Multimedia* 18(3):521–536. ISSN 1520-9210. <https://doi.org/10.1109/TMM.2016.2518485>

Load Balancing in Data Center Networks

Fei Li and Long Zhao
BUPT, Beijing, China

Synonyms

DCN traffic scheduling

Definitions

Load balancing uses advanced multilayer switching technology to provide real-time monitoring of server performance and health (Zhang et al. 2018). According to the health of different servers, the data traffic of the visitor is allocated to the appropriate server in the most economical and efficient manner. In this way, the load of each server can reach a relative balance, so that the resource sharing of the relevant data receiving device is optimal, the response time is short, the processor runs faster, the data exchange amount is greatly improved, and the operation efficiency is qualitatively leap.

Load balancing of data center networks usually involves two application scenarios. One is to evenly allocate large-scale concurrent access or network traffic to multiple server nodes to implement multiple servers responding to requests from different users. Another is to allocate the large-scale computational unit load evenly to multiple server nodes for arithmetic processing and then aggregate the operation nodes to feed back the user results.

Historical Background

With the rapid development of Internet technology and the widespread use of communication devices, the number of Internet users has increased rapidly, and the amount of data on the network has also increased dramatically. It can be said that data has entered an era of explosive growth. Therefore, new technologies are needed to store and process such large-scale data, and data center networks are put forward in this context.

The data center network refers to the network infrastructure of the data center, which connects a large number of servers through high-speed links and switches (Guo et al. 2008). The traditional data center network uses multiple tree topologies and consists of two or three layers of switch fabrics that can accommodate thousands of servers. However, in recent years, emerging

information services require a large amount of communication between virtual machines and servers, resulting in a sharp increase in Internal traffic in the data center and also presenting a new feature that is different from Internet traffic. Traditional tree-based data center architectures exhibit some shortcomings such as poor scalability, low utilization of network resources, and fragmentation of resources.

Therefore, academia and industry have begun to design new types of data center networks structures. Currently, there are two main types. One is a switch-centric structure and switches are used to exchange forwarding data, such as fat-tree (Greenberg et al. 2008) and Monsoon (Mysore et al. 2009). The other is a server-centric structure that uses servers to exchange forwarding data, such as BCube (Guo et al. 2009), DCeli, MDCube (Wu et al. 2009), etc. They make up for the shortcomings of the traditional tree-based data center networks structure from different aspects. Among them, the fat-tree topology is relatively simple and easy to deploy. At the same time, it also improves the problems of poor scalability, low utilization of network resources, and excessive load on the core layer switches in the traditional tree-based structure.

Foundations

In order to fully utilize the bandwidth resources in the network, the data center should employ the corresponding load balancing technology. By adaptively allocating the traffic among nodes throughout the network, the bandwidth utilization on different paths will be dynamically balanced in order to maximize the throughput of long data streams.

The traditional data center network adopts ECMP (equal-cost multipath) technology as the load balancing mechanism. Generally, the polling algorithm, the greedy algorithm, and the hash algorithm are used to evenly allocate the traffic load to multiple equal paths. That is, there are multiple paths between the source switch and the

destination switch to transmit data packets. The source switch distributes the traffic destined for the destination switch to multiple paths according to a load balancing algorithm, so that each path carries a part of the traffic and each path can be fully utilized. Therefore, it can avoid the use of only one path, which causes the network congestion, packet loss, and network throughput drop. However, due to its simplicity, the performance is greatly constrained.

How to optimally distribute ECMP traffic has always been a hot topic in academic fields. In order to solve the shortcomings of traditional ECMP, many adaptive dynamic load balancing techniques have been proposed. These technologies can be divided into three categories according to different mechanisms: centralized load balancing technology, switching device-based load balancing technology, and host-based load balancing technology.

Centralized Load Balancing Technology

Centralized load balancing technology relies on a centralized controller or host to specify the transmission path of the data stream. Typical centralized technologies, such as Hedera and Mahout, all take advantage of the idea of software-defined networking (SDN). They separate the pathfinding strategy and mechanism of the data stream, control the path of the data stream through centralized node control, and then dynamically update the routing table to achieve dynamic load balancing.

The centralized load balancing technology can effectively collect the link states of the entire network and determine the globally optimal load scheduling policy through calculation. At the same time, the centralized solution can also effectively deal with the fault paths that often occur in the network, making the scheduling strategy highly robust. However, the data center short data stream accounts for more than 80% of the total number of all data streams. The centralized load balancing technology is difficult to handle a large number of short data streams, so it is inevitable to face scalability and single point of failure problems. The biggest bottleneck of the load balancing technology is that the interaction

between the controller and the node takes a lot of time, which limits its development.

Switching Device-Based Load Balancing Technology

Unlike centralized load balancing technology, load balancing based on switching devices deploys the entire traffic scheduling function on switching devices in the network. Typical representations of such techniques are PacketScatter, LocalFlow, DRB, and CONGA.

Host-Based Load Balancing Technology

In order to make up for the shortcomings of centralized technology, many scholars have proposed a data center networks load balancing technology based on the host. This type of technology can operate without any centralized control. It uses the terminal system to detect the transmission status and link status in the network, calculates the appropriate routing path in the terminal system, and implements the scheduling control of the load through the source routing mechanism. Typical host-based load balancing technology includes MPTCP, FlowBender, and Presto.

Key Applications

It can be seen from the load balancing algorithms of the existing data center networks that the network architecture is very different in different scenarios, which also leaves a lot of redesigned space for the new scenarios. There are many new technologies, such as machine learning, that are deployed in data centers, and the emergence of these technologies may change the design of load balancing mechanisms.

Cross-References

► [Data Center Network Virtualization](#)

References

- Greenberg AG, Lahiri P, Maltz DA et al (2008) Towards a next generation data center architecture: scalability and commoditization. In: Proceedings of ACM PRESTO conference, Seattle, pp 57–62
- Guo C, Wu H, Tan K et al (2008) Dcell: a scalable and fault-tolerant network structure for data centers. In: Proceedings of ACM SIGCOMM conference, Seattle, pp 75–86
- Guo C, Lu G, Li D et al (2009) Bcube: a high performance, server-centric network architecture for modular data centers. In: Proceedings of ACM SIGCOMM conference, Barcelona, pp 63–74
- Mysore RN, Pamboris A, Farrington N et al (2009) Portland: a scalable fault-tolerant layer 2 data center network fabric. In: Proceedings of ACM SIGCOMM conference, Barcelona, pp 39–50
- Wu H, Lu G, Li D et al (2009) Mdcube: a high performance network structure for modular data center interconnection. In: Proceedings of ACM CoNEXT conference, Rome, pp 25–36
- Zhang J, Yu FR, Wang S et al (2018) Load balancing in data center networks: a survey. *IEEE Commun Surv Tutorials* 20(3):2324–2352

Local Cloud

► Data-Driven Edge Computing

Local Mobility Anchor

► Proxy Mobile IPv6

Localization in 3D Surface Wireless Sensor Networks

Miao Jin¹ and Hongyi Wu²

¹The Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, LA, USA

²The Center for Cybersecurity, Old Dominion University, Norfolk, VA, USA

Synonyms

Localization of wireless sensor networks deployed over complex 3D terrains

Problem Definition

Sensor network localization refers to the process of estimating the locations of sensor nodes with information between neighboring sensor nodes such as connectivity, local distance, and angle measurements. In real-world applications, many large-scale wireless sensor networks (WSNs) are deployed over complex terrains. Formally, a three-dimensional (3D) surface WSN is defined as follows.

Definition 1 (3D Surface Wireless Sensor Network) A 3D surface sensor network consists of sensor nodes deployed on a 3D surface where wireless signals between nearby nodes propagate along the surface only.

Definition 2 (Localization of 3D Surface Wireless Sensor Network) Given a 3D surface sensor network with distance measurements between neighboring nodes within their communication range, the localization problem is to recover the 3D coordinates of each sensor node.

Historical Background

Localization of a network deployed over a 3D surface generates a unique hardness compared with the well-studied localization of a network on 2D plane or in 3D volume. Specifically, due to limited radio range, the distance between two remote sensors deployed over a 3D surface can only be approximated by their surface distance, the length of the shortest path between them on the surface. Such surface distance is different from the 3D Euclidean distance of two nodes. A 3D surface is *localizable* if it exists a unique embedding up to a global rigid motion, under given constraints. Otherwise it is *non-localizable*. The following theorem claims that a network deployed over a 3D surface with surface distance information only is non-localizable, even if we assume accurate range distance measurement available (Zhao et al. 2012).

Theorem 1 *A general 3D surface is not localizable, given surface distance constraints only.*

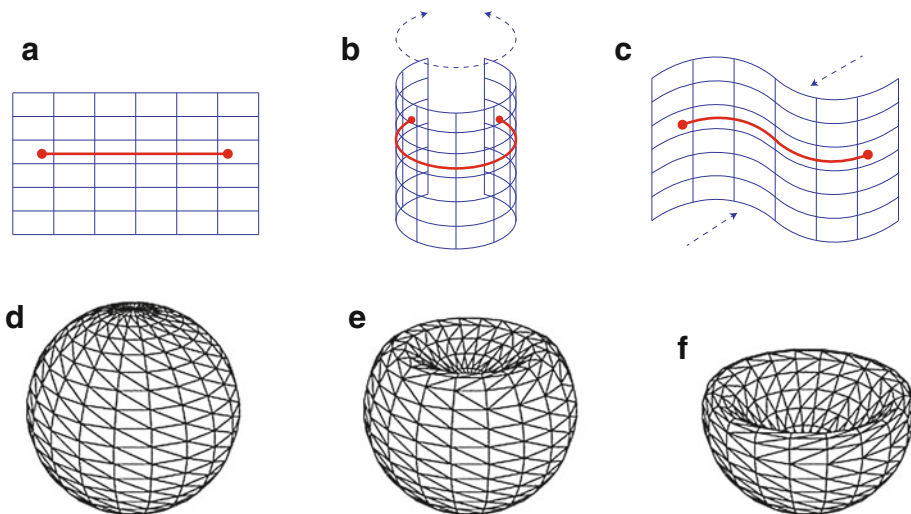
One intuitive example is that a piece of paper shown in Fig. 1a can be rolled up to a cylinder illustrated in Fig. 1b or to other curved shapes (see Fig. 1c). The distances between all pairs of points on the paper clearly remain the same under different shapes. Besides the open surface discussed above, similar deformation can be applied on closed surfaces (see Fig. 1d–f for example). Clearly, ambiguous embeddings exist, and thus a general 3D surface is not localizable under given surface distance constraints only.

Given a wireless sensor network deployed on a 3D terrain surface with one-hop distance information available, a simple distributed algorithm introduced in Zhou et al. (2011) extracts a refined triangular mesh from the network connectivity graph. Vertices of the triangular mesh represent the set of sensor nodes. An edge between two neighboring vertices indicates the communication link between the two sensors. The state-of-the-art surface network localization methods are based on the extracted triangular mesh structure.

Surface Network Localization with Height Information

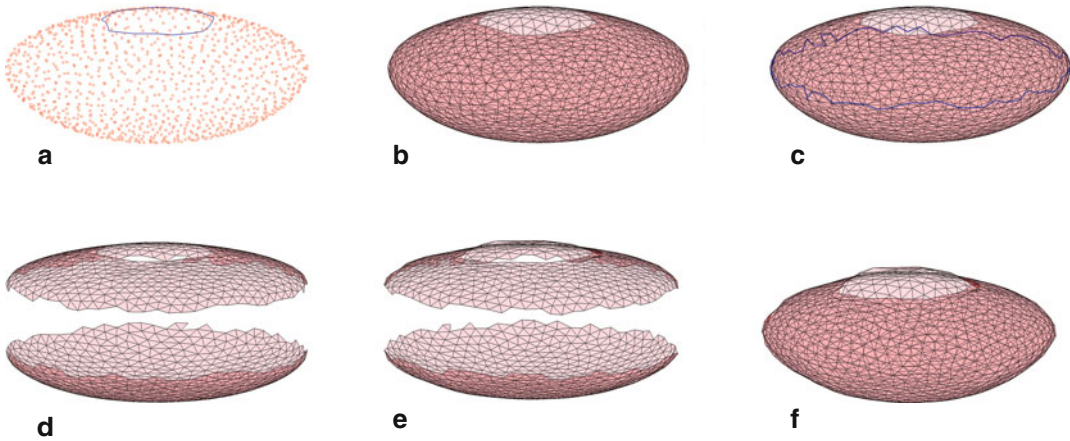
Under a practical setting with estimated link distances (between nearby nodes) and nodal heights (i.e., Z-coordinates) obtained by measuring atmospheric pressure, a sensor network deployed on a single-value 3D surface is localizable. Briefly, a single-value surface is one on which any two points have different projections on the X-Y plane. The definition is in reference to X-Y plane since sensors' heights are given as Z-coordinates. A planar projection of network deployed on a single-value surface converts a surface network localization problem to a planar one. The X-Y coordinates of nodes can be computed with well-known planar localization algorithms (Shang et al. 2003; Shang and Ruml 2004; Vivekanandan and Wong 2006; Lim and Hou 2009; Jin et al. 2011). The localization result is then mapped back to 3D by adding the known Z-coordinates.

For sensor networks deployed on general 3D surfaces, a distributed localization algorithm, dubbed cut-and-sew, applies a divide-and-



Localization in 3D Surface Wireless Sensor Networks, Fig. 1 Illustration of the non-localizability of general 3D surfaces. A surface can be deformed to another one without changing the surface distance between any pair

of points. (a) A 2D surface. (b) Deform to cylinder. (c) Deform to wave shape. (d) A closed 3D surface. (e) Small deformation. (f) Large deformation (Zhao et al. 2012)



Localization in 3D Surface Wireless Sensor Networks, Fig. 2 An overview of the cut-and-sew algorithm for general 3D surface localization. (a) A 3D surface network. (b) Triangular mesh. (c) Identified non-single-value edges.

(d) Partition of network to single-value patches. (e) Localization of individual patch. (f) Final localization result (Zhao et al. 2013)

conquer approach to localize sensor nodes with height information available. The basic idea is to partition a general 3D surface network into single-value patches, localizing individual patch and then merging them into a unified coordinate system. Note that the number of single-value patches should be minimized to avoid unnecessary partitioning and merging, which are subject to linear transformation errors. The key is to identify non-single-value edges that guide the division of a 3D surface network into single-value patches.

Figure 2 illustrates the basic idea of the cut-and-sew algorithm. Specifically, Fig. 2a, 2b shows a sensor network and the extracted triangular mesh, respectively. An edge is locally non-single-value if the projection of its two associated triangles overlaps on the X-Y plane. Each node in the network performs a simple and localized scheme to determine whether an associated edge is a non-single-value one based on its local distance information. Figure 2c shows the identified non-single-value edges. A set of non-single-value edges form the boundary of a single-value patch. Figure 2d illustrates the partition of network into a minimum set of single-value patches. Each single-value patch can then be localized as shown in Fig. 2e. Since neighboring patches share common vertexes and

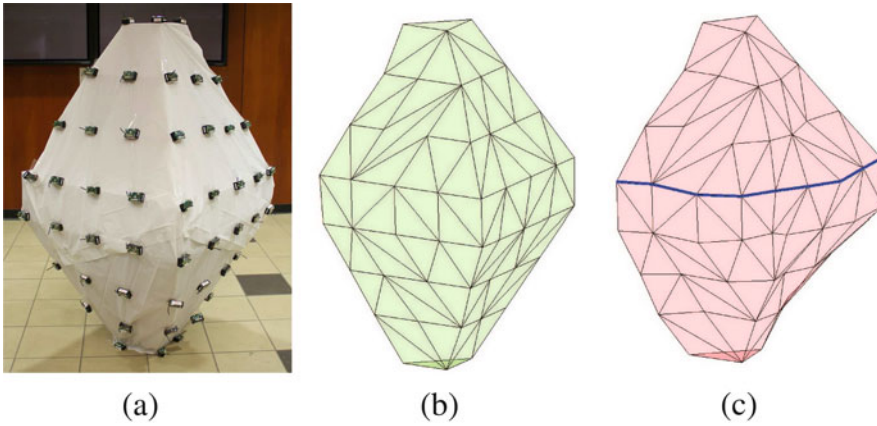
edges, they can be readily “sewed” together and yield a unified coordinate system for the entire 3D surface sensor network as illustrated in Fig. 2f.

Under practical sensor network settings, both surface distances and sensors heights are subject to measurement errors. The noisy distance and height measurements directly affect the identification of non-single-value edges, which may deviate from the ground truth and become isolated. One approach is to fuse nearby non-single-value edges to form a band and then cut the network along the medial axis of the band. This method effectively minimizes the impact of input errors on network partition and localization.

Prototyping and Experiments

Figure 3a shows one indoor test bed model built for prototyping 3D surface sensor networks with 5.25 feet high, 3 feet long, and 3 feet wide. Forty-eight Crossbow MICAz motes are attached to its surface. The algorithmic codes are implemented in Tiny OS and run on the Crossbow sensor nodes. The sensors are configured to use close to minimum radio transmission power (level 2) with a communication range between 25 and 55 cm. The short radio range avoids undesired connections through volume.

Every sensor periodically broadcasts a beacon message that contains its node ID to its



Localization in 3D Surface Wireless Sensor Networks, Fig. 3 Experimental setup and result. (a) Indoor 3D surface network test bed. (b) Input triangulation. (c)

Localization result (with an average location error of 14%) (Zhao et al. 2013)

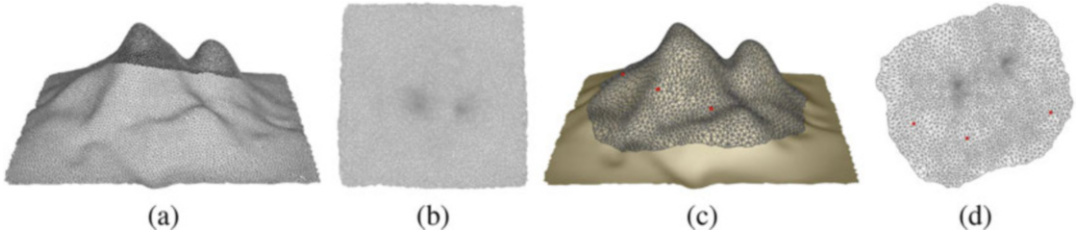
neighbors. Based on received beacon messages, a node builds a neighbor list with the RSSI of corresponding links. RSSI is used to estimate the length of links by looking up a RSSI-distance table established by experimental training data. The preliminary test shows that, under low transmission power, such estimation has an error rate about 20%. At the same time, the ground truth of surface distances and sensor coordinates is manually measured. Figure 3b illustrates the triangulation based on ground truth inputs. Figure 3c shows the localization result. The combined patches largely restore the original 3D surface network, with an average location error of about 14%.

Surface Network Localization with Digital Terrain Model

Integrating height measurement into every sensor of a network is not always practical and affordable, especially for a large-scale sensor network. However, a 3D representation of a terrain's surface, called digital terrain model (DTM), is available to public with a variable resolution up to one meter. DTMs are commonly built using remote sensing technology or from land surveying. A DTM is represented as a grid of squares, where the longitude, latitude, and altitude (i.e., 3D coordinates) of all grid points are known. It

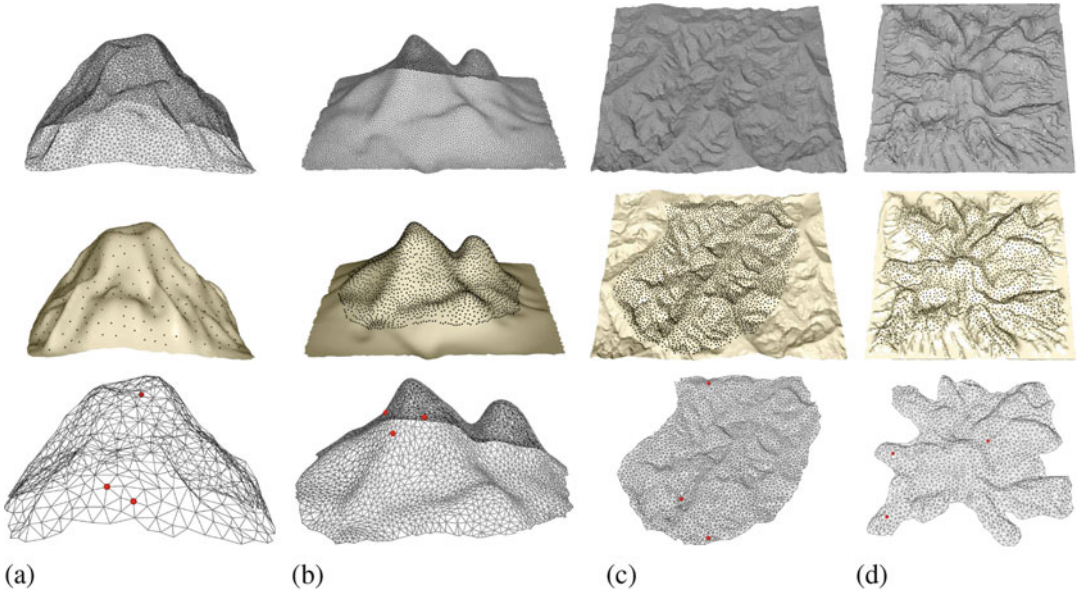
is straightforward to convert a grid into a triangulation, e.g., by simply connecting a diagonal of each square. Therefore a triangular mesh of a terrain surface can be available before we deploy a sensor network on it. On the other hand, a refined triangular mesh can be extracted from the connectivity graph of network deployed on the terrain surface with local distance information. The constraint that the sensors must be on the known 3D terrain surface ensures that the triangular meshes of terrain surface and network approximate the same geometric shape. Theoretically, the two triangular meshes share the same conformal structure. We can construct a well-aligned conformal mapping between them. Based on this mapping, each sensor node of the network can easily locate reference grid points of the DTM to calculate its own location.

Figure 4 illustrates the basic idea to localize a surface network on a 3D terrain surface with DTM available. Figure 4a, c shows the triangular meshes of a terrain surface and a network deployed over the terrain surface, respectively. We first compute two conformal mappings, denoted as f_1 and f_2 respectively, to map the two triangular meshes to plane as shown in Fig. 4b, d respectively. However, the two mapped triangular meshes on plane are not aligned. Three anchor nodes, sensor nodes equipped with GPS, marked with red as shown in Fig. 4c are deployed with the network to provide a reference for alignment.



Localization in 3D Surface Wireless Sensor Networks, Fig. 4 (a) The triangular mesh of a terrain surface. (b) The triangular mesh of the terrain surface is conformally mapped to plane. (c) The triangular mesh of a network

deployed on the terrain surface with three randomly deployed anchor nodes marked with red. (d) The triangular mesh of the network is conformally mapped to plane (Yang et al. 2014)



Localization in 3D Surface Wireless Sensor Networks, Fig. 5 The first row shows a set of DTMs of representative terrain surfaces. The second row shows sensor nodes marked with black points randomly deployed on these

terrain surfaces. The third row shows the localized sensor networks with anchor nodes marked with red (Yang et al. 2014)

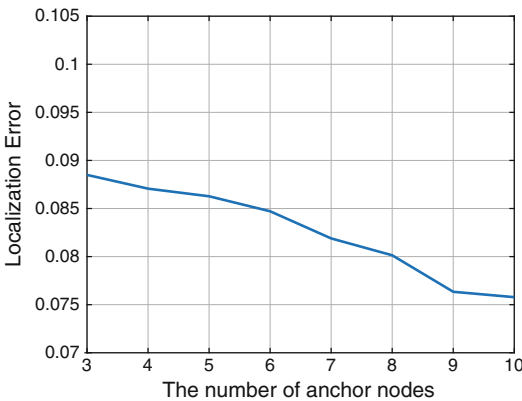
Based on the positions of the three anchor nodes, we construct a Möbius transformation, denoted as f_3 , to align the mapped triangular meshes of the network and the terrain surface on plane. Combining the three mappings, $f_1^{-1} \circ f_3 \circ f_2$, induces a well-aligned conformal mapping between the two triangular meshes shown in Fig. 4a, c, respectively. Based on the well-aligned mapping, each sensor node of the network simply locates its nearest grid points, vertices of the triangular mesh of the terrain surface, to calculate its own geographic location.

Deployment of Anchor Nodes

Figure 5 shows a set of representative terrain surfaces and their corresponding DTMs, on which sensor nodes marked as black points are randomly deployed. For each network, eight tests are repeated with a various deployment of three anchor nodes in each test. Table 1 shows the mean (μ), median ($\tilde{x}$), and standard deviation (σ) of localization errors under different deployments of anchor nodes of each network. The positions of three anchor nodes affect the performance of the localization algorithm slightly. In general,

Localization in 3D Surface Wireless Sensor Networks, Table 1 The distribution of localization errors under different sets of anchor nodes (Yang et al. 2014)

Error	DTM I	DTM II	DTM III	DTM IV
μ	0.2579	0.1356	0.0951	0.2098
$\tilde{x}$	0.2306	0.1343	0.0956	0.1512
σ	0.1089	0.1717	0.0158	0.0352



Localization in 3D Surface Wireless Sensor Networks, Fig. 6 Localization error decreases slightly with the increased number of anchor nodes (Yang et al. 2014)

the more scattered the three anchor nodes are deployed in a network, the lower the localization error is.

Size of Anchor Nodes

Theoretically speaking, the localization algorithm requires only three anchor nodes to align two triangular meshes on plane, regardless of the size of a network. If more anchor nodes are deployed on the network, a least-square conformal mapping method can replace Möbius transformation to incorporate all anchor nodes into the alignment to improve the localization accuracy. Figure 6 shows one example. The localization error of a network with size 2.6 k deployed on a 3D surface, shown in Fig. 4d, decreases slightly with the increased number of anchor nodes.

Applications

Geographic location information is imperative to a variety of applications in WSNs, ranging from position-aware sensing to geographic routing. While global navigation satellite systems (such as GPS) have been widely employed for localization, integrating a GPS receiver in every sensor of a large-scale sensor network is unrealistic due to high cost. Moreover, some application scenarios prohibit the reception of satellite signals by part or all of the sensors, rendering it impossible to solely rely on global navigation systems. In real-world applications, many large-scale WSNs are deployed over complex terrains, such as the volcano monitoring project (Werner-Allen et al. 2006). The introduced state-of-the-art surface network localization algorithms recover the coordinates of sensor nodes with assumption of the availability of nodal height measurements (Zhao et al. 2012, 2013) or digital terrain model, a 3D representation of terrain’s surface (Yang et al. 2014). They are all distributed and scalable to large-scale sensor networks deployed over general 3D terrain surfaces.

Cross-References

- ▶ Cooperative Localization in Wireless Sensor Networks
- ▶ Fingerprinting Localization
- ▶ QoS in Indoor Localization

References

Jin M, Xia S, Wu H, Gu X (2011) Scalable and fully distributed localization with mere connectivity. In: Proceeding of INFOCOM, Shanghai, pp 3164–3172

Lim H, Hou J (2009) Distributed localization for anisotropic sensor networks. *ACM Trans Sensor Netw* 5(2):11–37

Shang Y, Ruml W (2004) Improved MDS-based localization. In: Proceeding of INFOCOM, Hong Kong, pp 2640–2651

Shang Y, Ruml W, Zhang Y, Fromherz MPJ (2003) Localization from mere connectivity. In: Proceeding of MobiHoc, Annapolis, MD, pp 201–212

- Vivekanandan V, Wong VWS (2006) Ordinal MDS-based localization for wireless sensor networks. *Int J Sensor Netw* 1(3/4):169–178
- Werner-Allen G, Lorincz K, Johnson J, Lees J, Welsh M (2006) Fidelity and yield in a volcano monitoring sensor network. In: *Proceedings of the 7th symposium on operating systems design and implementation, OSDI06*, Seattle, WA, pp 381–396
- Yang Y, Jin M, Wu H (2014) 3D surface localization with terrain model. In: *Proceeding of IEEE conference on computer communications (INFOCOM)*, Toronto, pp 46–54
- Zhao Y, Wu H, Jin M, Xia S (2012) Localization in 3D surface sensor networks: challenges and solutions. In: *Proceeding of the 31st annual IEEE conference on computer communications (INFOCOM12)*, Orlando, FL, pp 55–63
- Zhao Y, Wu H, Jin M, Yang Y, Zhou H, Xia S (2013) Cut-and-sew: a distributed autonomous localization algorithm for 3D surface wireless sensor networks. In: *Proceeding of the 14th ACM international symposium on mobile ad hoc networking and computing (MobiHoc13)*, Bangalore
- Zhou H, Wu H, Xia S, Jin M, Ding N (2011) A distributed triangulation algorithm for wireless sensor networks on 2D and 3D surface. In: *Proceeding of INFOCOM*, Shanghai, pp 1053–1061

Localization of Wireless Sensor Networks Deployed over Complex 3D Terrains

► Localization in 3D Surface Wireless Sensor Networks

Location Based Service of Ad Hoc Networks

Shuai Han¹, Deyue Zou², and Weixiao Meng³

¹School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China

²Dalian University of Technology, Dalian, China

³Harbin Institute of Technology, Harbin, China

Synonyms

Positioning-Locate-Locating; Position-Locate; Radiomap-Database

Definitions

This entry provides information about the ad hoc net-based LBS, including the histories, the positioning principles, the classical strategies, and the key applications.

Historical Background

In 1997, Ward et al. (1997), Welch et al. (1999), and Priyantha et al. (2000) proposed localization systems rely on beacons. All beacon-based positioning algorithms, however, have their limitations because they need another positioning scheme to bootstrap the beacon node positions. Therefore, in 2004, beacon-free algorithms were proposed in Shang and Ruml (2004), Ji and Zha (2004), and Priyantha et al. (2003). Beacon-free algorithms utilize local distance information rather than beacons to estimate the relative coordinates of each node.

In 2002, a signal strength decay-based ranging scheme was implemented in Patwari et al. (2002). This work belongs to range-based localization. In contrast, range-free algorithms do not require distance estimation or angle estimation. For example, connectivity matrix of sensor nodes can be used for position estimation (Doherty et al. 2001); the Euclidian distance estimated by hop-count can be used to estimate the node locations via triangulation method (Niculescu and Nath 2003; Nagpal et al. 2003). Incremental algorithms are used in ad hoc network positioning. At the beginning, only three or four core nodes are located, and then more nodes are located based on the positions of the core nodes (Savarese et al. 2001). However, incremental algorithms have error propagation problem, resulting in poor overall performance. Concurrent strategies are invented to solve this problem, all nodes are approximately located at the beginning, and then the node positions would be optimized based on the node-to-node distances (Priyantha et al. 2003).

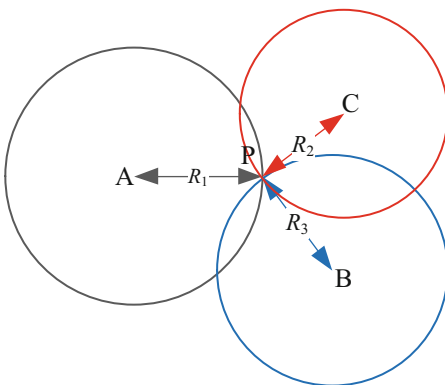
Basic Positioning Principles (Liu et al. 2007)

Triangulation

Triangulation uses the geometric properties of triangles to estimate the target location. It has two derivations: ranges and angulation. Range estimates the position of an object by measuring it from multiple reference points. So, it is called range measurement techniques. Instead of measuring the distance directly using received signal strengths (RSS), time of arrival (TOA) or time difference of arrival (TDOA) is usually measured, and the distance is derived by computing the attenuation of the emitted signal strength or by multiplying the radio signal velocity and the travel time. Roundtrip time of flight (RTOF) or received signal phase method is also used for range estimation in some systems. Angulation locates an object by computing angles relative to multiple reference points.

Ranging Techniques

- TOA: The distance from the mobile target to the measuring unit is directly proportional to the propagation time. In order to enable 2-D positioning, TOA measurements must be made with respect to signals from at least three reference points, as shown in Fig. 1 (Kim et al. 2016a).



Location Based Service of Ad Hoc Networks, Fig. 1
Positioning based on TOA/RTOF measurements

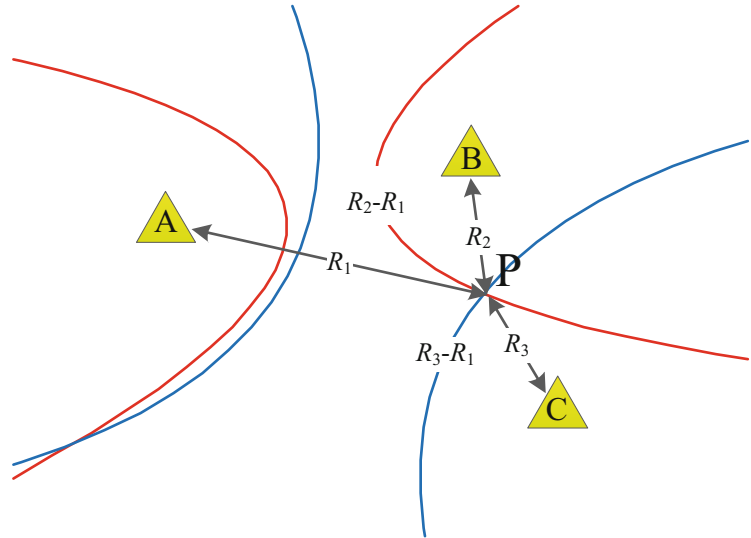
For TOA-based systems, the one-way propagation time is measured, and the distance between measuring unit and signal transmitter is calculated. In general, direct TOA results in two problems. First, all transmitters and receivers in the system have to be precisely synchronized. Second, a timestamp must be labeled in the transmitting signal in order for the measuring unit to discern the distance the signal has traveled. TOA can be measured using different signaling techniques such as direct sequence spread-spectrum (DSSS) (Liu et al. 2014; Ye et al. 2016) or Ultra Wide Band (UWB) measurements (Zadgaonkar and Chandak 2016). A straightforward approach uses a geometric method to compute the intersection points of the circles of TOA. The position of the target can also be computed by minimizing the sum of squares of a nonlinear cost function, i.e., least-squares algorithm (Kim et al. 2016a; Gu et al. 2017). There are other algorithms for TOA-based location system such as closest-neighbor (CN) and residual weighting (RWGH) (Gu et al. 2017). The CN algorithm estimates the location of the user as the location of the base station or reference point that is located closest to that user. The RWGH algorithm can be basically viewed as a form of weighted least-squares algorithm. It is suitable for LOS, non-LOS (NLOS), and mixed LOS/NLOS channel conditions.

- TDOA: The idea of TDOA is to determine the relative position of the mobile transmitter by examining the difference in time at which the signal arrives at multiple measuring units, rather than the absolute arrival time of TOA. For each TDOA measurement, the transmitter must lie on a hyperboloid with a constant range difference between the two measuring units. A 2-D target location can be estimated from the two intersections of two or more TDOA measurements, as shown in Fig. 2.

Two hyperbolas are formed from TDOA measurements at three fixed measuring units (A, B, and C) to provide an intersection point, which locates the target P. The conventional methods for computing TDOA estimates are

Location Based Service of Ad Hoc Networks,

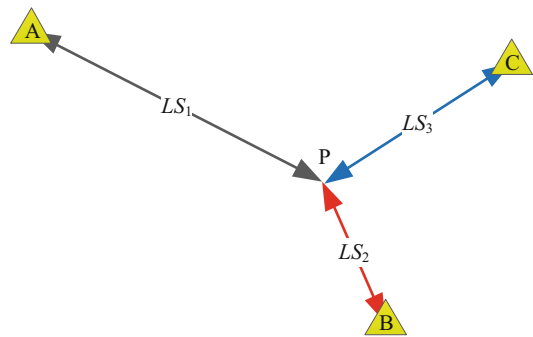
Fig. 2 Positioning based on TDOA measurements



to use correlation techniques. TDOA can be estimated from the cross correlation between the signals received at a pair of measuring units. The above two schemes have some drawbacks. For indoor environments, it is difficult to find a LOS channel between the transmitter and the receiver. Radio propagation in such environments would suffer from multipath effect. The time and angle of an arrival signal would be affected by the multipath effect; thus, the accuracy of estimated location could be decreased.

- **RSS-Based (or Signal Attenuation-Based) Method:** The power, or energy, of a signal traveling between two nodes is a signal parameter that contains information related to the distance (“range”) between those nodes. This parameter, commonly referred to as RSS, can be used together with a path loss and shadowing model to provide a distance estimate (Gezici 2008). Signal attenuation-based methods attempt to calculate the signal path loss due to propagation. Theoretical and empirical models are used to translate the difference between the transmitted signal strength and the received signal strength into a range estimate, as shown in Fig. 3.

Due to severe multipath fading and shadowing present in the indoor environment, path loss models do not always hold. The



Location Based Service of Ad Hoc Networks, Fig. 3 Positioning based on RSS, where LS1, LS2, and LS3 denote the measured path loss

parameters employed in these models are site-specific. The accuracy of this method can be improved by utilizing the premeasured RSS contours centered at the receiver (Van et al. 2017) or multiple measurements at several base stations. A fuzzy logic algorithm shown in Fallah and Pourahmadi (2018) is able to significantly improve the location accuracy using RSS measurement.

- **RTOF:** This method is to measure the time-of-flight of the signal traveling from the transmitter to the measuring unit and back, called the RTOF (see Fig. 1). For RTOF, a more moderate relative clock synchronization requirement replaces the above synchronization require-

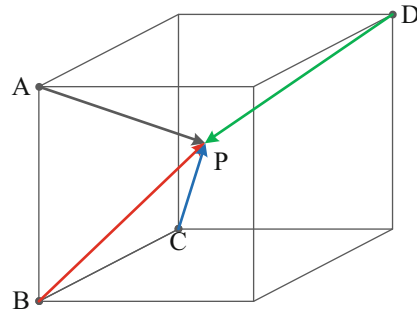
ment in TOA. Its range measurement mechanism is the same as that of the TOA. The measuring unit is considered as a common radar. A target transponder responds to the interrogating radar signal, and the complete roundtrip propagation time is measured by the measuring units. However, it is still difficult for the measuring unit to know the exact delay/processing time caused by the responder in this case. In long-range or medium-range systems, this delay could be ignored if it is small, compared with the transmission time. However, for short-range systems, it cannot be ignored. An alternative approach is to use the concept of modulated reflection (Kim et al. 2016b), which is only suited for short-range systems. An algorithm to measure RTOF of wireless LAN packets is presented in D'Souza et al. (2016) with the result of a measurement error of a few meters. The positioning algorithms for TOA can be directly applicable for RTOF.

- **Received Signal Phase Method:** The received signal phase method uses the carrier phase (or phase difference) to estimate the range. This method is also called phase of arrival (POA) (Makki et al. 2015). Assuming that all transmitting stations emit pure sinusoidal signals that are of the same frequency, with zero phase offset, in order to determine the phases of signals received at a target point, the signal transmitted from each transmitter to the receiver needs a finite transit delay. In Fig. 4, the transmitter stations A up to D are placed at particular locations within an imaginary cubic building.

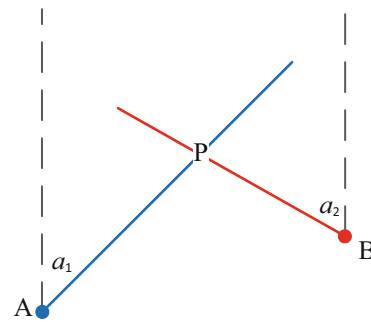
The receiver may measure phase differences between two signals transmitted by pairs of stations, and positioning systems are able to adopt the algorithms using TDOA measurement to locate the target.

Angulation Techniques (AOA Estimation)

In AOA, the location of the desired target can be found by the intersection of several pairs of angle direction lines, each formed by the circular radius from a base station or a beacon station to the



Location Based Service of Ad Hoc Networks, Fig. 4
Positioning based on signal phase



Location Based Service of Ad Hoc Networks, Fig. 5
Positioning based on AOA measurement

mobile target. As shown in Fig. 5, AOA methods may use at least two known reference points (1, 2) and two measured angles a_1 and a_2 to derive the 2-D location of the target P. Estimation of AOA, commonly referred to as direction finding (DF), can be accomplished either with directional antennae or with an array of antennae.

The advantages of AOA are that a position estimate may be determined with as few as three measuring units for 3-D positioning or two measuring units for 2-D positioning and that no time synchronization between measuring units is required. The disadvantages include relatively large and complex hardware requirement(s) and location estimate degradation as the mobile target moves farther from the measuring units. For accurate positioning, the angle measurements need to be accurate, but the high accuracy measurements in wireless networks may be limited by shadowing, by multipath reflections arriving from misleading directions, or by the directivity

of the measuring aperture. Some literatures also call AOA as direction of arrival (DOA). For more detailed discussions on AOA estimation algorithms and their properties, see Wu et al. (2017), Gui et al. (2017), and Gu and Ren (2015).

Scene Analysis

For triangulation methods, line-of-sight (LOS) scenario is always necessary, but in non-line-of-sight (NLOS) scenarios, scene analysis methods are more trustworthy.

RF-based scene analysis refers to the type of algorithms that first collect features (fingerprints) of a scene and then estimate the location of an object by matching online measurements with the closest *a priori* location fingerprints.

The term “location fingerprinting” covers a wide variety of methods for determining receiver position using databases from different sources, such as RSS, TOA, TDOA, multi-pass feature, and so on (Honkavirta et al. 2009). RSS-based location fingerprinting is commonly used in scene analysis. Location fingerprinting refers to techniques that match the fingerprint of some characteristic of a signal that is location dependent.

There are two stages for location fingerprinting: offline stage and online stage (or run-time stage). During the offline stage, a site survey is performed in an environment. The location coordinates/labels and respective signal strengths from nearby base stations/measuring units are collected. During the online stage, a location positioning technique uses the currently observed signal strengths and previously collected information to figure out an estimated location. There are at least five location fingerprinting-based positioning algorithms using pattern recognition technique so far: probabilistic methods, *k*-nearest-neighbor (kNN), neural networks, support vector machine (SVM), and smallest *M*-vertex polygon (SMP).

The techniques based on location fingerprinting utilize the diffraction, reflection, and scattering in the propagation indoor environments. The main challenge to fingerprinting-based technology is that the database is time-consuming, labor-intensive, and vulnerable to environmental

changes and the process requires certain pedigree on the surveyor that may deem the fingerprinting techniques impractical to be deployed over large areas (e.g., shopping malls, multistory offices/residences, etc.) (Hossain and Soh 2015). Researchers are trying to find some fast-rate fingerprint collection methods (Schmidt et al. 2018), such as crowdsourcing methods (Wang et al. 2016; Zou et al. 2013) and radio map simulation methods.

Proximity

Proximity algorithms provide symbolic relative location information. Usually, it relies upon a dense grid of antennas, each having a well-known position. When a mobile target is detected by a single antenna, it is considered to be collocated with it. When more than one antenna detects the mobile target, it is considered to be collocated with the one that receives the strongest signal. This method is relatively simple to implement. It can be implemented over different types of physical media. In particular, the systems using infrared radiation (IR) and radio frequency identification (RFID) are often based on this method. Another example is the Cell Identification (Cell-ID) or Cell of Origin (COO) method. Furthermore, many of the wireless sensor networks (WSN)-based positioning systems use proximity algorithms (Wang et al. 2010).

Positioning Strategies for Ad Hoc Sensor Networks Sun et al. (2005)

The unique properties of sensor networks cause them to require somewhat different positioning strategies than those used in cellular networks and WLANs. There are various possibilities for constructing a localization strategy that balances computation among sensor network nodes. These strategy designs include the distributed (every node should be able to estimate its own location), the localized (each node gathers information from other nodes in its immediate neighborhood), the asymptotic convergence design (computation stops when a certain degree of accuracy has

been achieved), the self-organizing scheme (node functioning does not depend on the global infrastructure), the robust design (the strategy can tolerate node failures and range errors), and the cost-effective and energy-efficient approach (this strategy requires little computation overhead).

Beacon-Based Versus Beacon-Free Strategies

Beacon-based strategies assume that a certain minimum number of nodes know their own positions through manual configuration or GPS. Individual nodes' location could then be determined by referring to a beacon's position. A detailed survey of various beacon-based localization strategies has been provided in Bulusu et al. (2004). All beacon-based positioning strategies, however, have their limitations because they need another positioning scheme to bootstrap the beacon node positions, and they cannot be easily employed in environments where other location systems are unavailable; thus, they are unsuitable for use indoors. It turns out, in practice, that a large number of beacon nodes are required to achieve an acceptable level of position error (Bulusu et al. 2004). In contrast, beacon-free strategies use local distance information to attempt to determine each node's relative coordinates without relying on beacons that are aware of their positions. Of course, any strategy that does not use beacon nodes can be easily converted to one that uses a small number of beacon nodes by adding a final step in the procedure, in which all node coordinates are transformed using three (in 2-D positioning) or four (in 3-D positioning) beacon nodes. For example, multidimensional scaling (MDS) has been investigated for use in solving the localization problem in sensor networks (Shang and Ruml 2004; Ji and Zha 2004). Without using beacons, all MDS-based localization strategies are able to produce maps that represent the relative positions of nodes. With three or more beacons, the absolute coordinates of all the nodes can be determined at the same time. Another fully distributed and beacon-free localization strategy, proposed in Priyantha et al. (2003), operates in two stages.

In the first stage, a heuristic is employed to create a well-spread, fold-free graph layout that resembles the desired layout. The second stage uses a mass-spring model analog. The optimized localization estimates analog can be found to be the minimum energy stage of the mass-spring model.

Incremental Versus Concurrent Strategies

Incremental strategies begin with only three or four core nodes being aware of their own coordinates. They then recursively add appropriate nodes to this set by calculating each node's coordinate using measured relative distances from nodes with previously known coordinates. These coordinate calculations are based on either simple triangle strategies or on local optimization schemes (Savarese et al. 2001). However, incremental strategies have limitations because they can propagate measurement error, resulting in poor overall network localization. Some incremental strategies can thus be applied in later stages of global optimization to reduce propagation error. Escaping from local minima and reaching global minima in the incremental stage continues to be a major challenge. In concurrent strategies, all nodes calculate and then refine their coordinates in parallel using local information. Some of these strategies use iterative optimization schemes that reduce the differences between measured distances and calculated distances based on current coordinate estimates (Priyantha et al. 2003). Concurrent optimization strategies have a better chance of avoiding local minima than incremental schemes. They can also avoid error propagation by continuously reducing global errors.

Range-Based Versus Range-Free Strategies

Many localization strategies rely on the distances between nodes: this is known as range-based localization. This type of distance estimation is usually implemented using signal strength decay, TOA, or TDOA for internode range estimation (Patwari et al. 2002). While range-based strategies require absolute point-to-point distance estimation (range) or angle estimation for positioning, range-free strategies

do not require this information. In addition to measuring range information, range-free localization strategies achieve position estimation by solving a convex optimization problem using a connectivity matrix of sensor nodes (Doherty et al. 2001). However, due to the unique ad hoc character of wireless sensor networks, many distributed solutions are more attractive than centralized designs. Sensors locate themselves at the centroid of the locations of beacons they can detect (Bulusu et al. 2000). Other strategies have assumed node-to-node communication can be used to convey the locations of all the beacons to all the sensors (Niculescu and Nath 2003; Nagpal et al. 2003). Using hop-count as an estimate of Euclidian distance (by computing the average distance between sensors in Niculescu and Nath (2003) or by analytically deriving it in Nagpal et al. 2003), a sensor can estimate its position via triangulation. Another novel “range-free” strategy determines whether a node lies inside or outside of the triangles formed by all possible sets of three beacons (called the APIT test) (He et al. 2003).

Single-Hop Versus Multiple-Hop Strategies

In multi-hop positioning systems, nodes typically do not receive beacon nodes’ signals directly. Given the influence of single-hop positioning systems such as cellular networks and WLANs, it is not surprising that the first multi-hop localization strategies tried to adapt single-hop technologies. TDOA, RSSI, and/or AOA information is collected, and the position of each node is then computed using triangulation (Niculescu and Nath 2003). A number of connectivity-based solutions have been proposed for multi-hop localization. One of the simplest and earliest is DV-hop (Niculescu and Nath 2003). In this system, each node determines its own position based on how many hops away it is from a beacon node. A method similar to APS was suggested in Nagpal et al. (2003). It first determines the hop distance (called gradient) to the beacons (called seeds) and, as a function of the average node density, calculates the actual average hop distance to a beacon.

Key Applications

Routing Optimize

For most of the self-organizing network technologies, position information is needed for relaying (Mauve et al. 2001; Liu et al. 2016). Routing problems of self-organizing network will be solved more convenient with the help of location information, because the distribution of the nodes is known and the power cost and the time delay of the multi-hop communication would be optimized (Li et al. 2009; Noguchi and Kobayashi 2017; Pham Thi Minh et al. 2015).

Node Placement

The working efficiency of ad hoc network is affected by the distribution of the nodes (Erdogan et al. 2003; Yan and Olariu 2009). To optimize this distribution of the nodes, the location information of the node is necessary. LBS provides this position information, and the nodes can be placed at right places or automatically move to the optimized place.

Target Detecting

Ad hoc net-based LBS can also be used for target detecting. The location of cooperative user can be calculated by the ad hoc net by ranging-based algorithms. And the position of uncooperative user can be detected by the RSS changes; this method can be used in security systems, landmine systems, or antisubmarine systems.

BAN Applications

Body area network (BAN) organizes the nearby gears of a user to provide a personalized service. Such a people-centered network needs the location of the user to organize the network, which can be provided by the LBS of ad hoc network.

IoT Applications

Internet of Things (IoT) is playing an important role in the society, i.e., express, library management, and find things. These applications all depend on the location information of things. Also, the network of these things are ad hoc networks, so ad hoc-based LBS becomes the first choice for the positioning services of them.

Citations

- Journal article: Ward et al. (1997), Welch et al. (1999), Patwari et al. (2002), Liu et al. (2007, 2014, 2016), Kim et al. (2016a,b), Ye et al. (2016), Zadgaonkar and Chandak (2016), Gu et al. (2017), Gezici (2008), Van et al. (2017), Fallah and Pourahmadi (2018), D'Souza et al. (2016), Makki et al. (2015), Wu et al. (2017), Gui et al. (2017), Gu and Ren (2015), Hossain and Soh (2015), Schmidt et al. (2018), Wang et al. (2010, 2016), Sun et al. (2005), Bulusu et al. (2000), and Mauve et al. (2001)
- Conference article: Priyantha et al. (2000, 2003), Shang and Ruml (2004), Ji and Zha (2004), Doherty et al. (2001), Niculescu and Nath (2003), Nagpal et al. (2003), Savarese et al. (2001), Honkavirta et al. (2009), Zou et al. (2013), Bulusu et al. (2004), He et al. (2003), Li et al. (2009), Noguchi and Kobayashi (2017), Pham Thi Minh et al. (2015), Erdogan et al. (2003), and Yan and Olariu (2009).

References

- Bulusu N, Heidemann J, Estrin D (2000) GPS-less low cost outdoor localization for very small devices. *IEEE Pers Commun* (Special Issue on Smart Spaces and Environments) 7(5):28–34
- Bulusu N, Heidemann J, Estrin D, Tran T (2004) Self-configuring localization systems: design and experimental evaluation. *ACM Trans Embed Comput Syst* 3(1):24–60
- Doherty L, Pister K, Ghaoui L (2001) Convex position estimation in wireless sensor networks. In: *Proceedings of IEEE Infocom 2001*, Anchorage, vol 3, 22–26 Apr 2001, pp 1655–1663
- D'Souza M, Schoots B, Ros M (2016) Indoor position tracking using received signal strength-based fingerprint context aware partitioning. *Radar Sonar Navig IET* 10(8):1347–1355
- Erdogan A, Cayirci E, Coskun V (2003) Sectoral sweepers for sensor node management and location estimation in adhoc sensor networks. In: *IEEE military communications conference, MILCOM'03*, Boston, vol 1, pp 555–560
- Fallah M, Pourahmadi V (2018) Graph-based iterative measurement – denoising and radio-map generation for semi-supervised indoor localisation. *Commun IET* 12(7):848–853
- Gezici S (2008) A survey on wireless position estimation. *Wirel Pers Commun* 44(3):263–282
- Gu Y, Ren F (2015) Energy-efficient indoor localization of smart hand-held devices using bluetooth. *Access IEEE* 3:1450–1461
- Gu X, Cui Y, Wang Q, Yuan H, Zhao L, Wu G (2017) Received signal strength indication-based localization method with unknown path-loss exponent for HVDC electric field measurement. *High Voltage* 2(4): 261–266
- Gui L, Yang M, Fang P, Yang S (2017) RSS-based indoor localization using MDCF. *Wirel Sens Syst IET* 7(4):98–104
- He T, Huang C, Blum B, Stankovic JA, Abdelzaher T (2003) Range-free localization schemes in large scale sensor networks. In: *Proceedings of 9th annual ACM international conference on mobile computing and networking (MobiCom)*, Sept 2003, pp 81–95
- Honkavirta V, Perala T, Ali-Loytty S et al (2009) A comparative survey of WLAN location fingerprinting methods. *Workshop on positioning*. IEEE
- Hossain AKMM, Soh WS (2015) A survey of calibration-free indoor positioning systems. *Comput Commun* 66:1–13
- Ji X, Zha H (2004) Sensor positioning in wireless ad-hoc sensor networks with multidimensional scaling. In: *Proceedings IEEE Infocom 2004*, pp 2652–2661
- Kim NY, Yoon S-W, Kim C-W (2016a) Cross-connection abatement in dense loosely coupled wireless charging pad environments. *Electron Lett* 52(7):553–555
- Kim J-Y, Yang S-H, Son Y-H, Han S-K (2016b) High-resolution indoor positioning using light emitting diode visible light and camera image sensor. *Optoelectron IET* 10(5):184–192
- Li Z, Zhu Y, Li M (2009) Practical location-based routing in vehicular ad hoc networks. In: *2009 IEEE 6th international conference on mobile adhoc and sensor systems*, Macau, pp 900–905
- Liu H, Darabi H, Banerjee P, Liu J (2007) Survey of wireless indoor positioning techniques and systems. *IEEE Trans Syst Man Cybern* 37(6):1067–1080
- Liu Q, Linton D, Fusco V (2014) Compact CRLH sensor for target localization. *Microw Opt Technol Lett* 56:1883
- Liu J, Wan J, Wang Q et al (2016) A survey on position-based routing for vehicular ad hoc networks. *Telecommun Syst* 62(1):15–30
- Makki A, Siddig A, Saad MM, Bleakley CJ, Cavallaro JR (2015) High-resolution time of arrival estimation for OFDM-based transceivers. *Electron Lett* 51(3):294–296
- Mauve M, Widmer A, Hartenstein H (2001) A survey on position-based routing in mobile ad hoc networks. *IEEE Netw* 15(6):30–39
- Nagpal R, Shrobe H, Bachrach J (2003) Organizing a global coordinate system from local information on an ad hoc sensor network. In: *Proceedings of 2nd international workshop information processing in sensor networks (IPSN'03)*, Apr 2003, pp 333–348

- Niculescu D, Nath B (2003) Ad hoc positioning system (APS) using AOA. In: *Proceedings IEEE INFOCOM 2003*, San Francisco, vol 3, pp 1734–1743
- Noguchi T, Kobayashi T (2017) Adaptive location-aware routing with directional antennas in mobile adhoc networks. In: *2017 international conference on computing, networking and communications (ICNC)*, Santa Clara, pp 1006–1011
- Patwari N, Hero III AO, Perkins M, Correal NS, O'Dea RJ (2002) Relative location estimation in wireless sensor networks. *IEEE Trans Signal Process (Special Issue on Signal Processing in Networks)*, 51(8):2137–2148
- Pham Thi Minh T, Nguyen TT, Kim D (2015) Location aided zone routing protocol in mobile ad hoc networks. In: *2015 IEEE 20th conference on emerging technologies & factory automation (ETFA)*, Luxembourg, pp 1–4
- Priyantha N, Chakraborty A, Balakrishnan H (2000) The cricket location support system. In: *Proceedings of ACM MobiCom 2000*. ACM, Boston, pp 32–43
- Priyantha NB, Balakrishnan H, Teller S (2003) Anchor-free distributed localization in sensor networks. In: *Proceedings of 1st international conference embedded networked sensor systems (SenSys 2003)*, Los Angeles, 5–7 Nov 2003, pp 340–341
- Savarese C, Rabaey JM, Beutel J (2001) Location in distributed ad-hoc wireless sensor networks. In: *Proceedings of international conference on acoustics, speech and signal processing (ICASSP'01)*, pp 2037–2040
- Schmidt E, Mohammed MA, Akopian D (2018) A performance study of a fast-rate WLAN fingerprint measurement collection method. *IEEE Trans Instrum Meas*, Oct 2018, 67(10):2273–2281
- Shang Y, Ruml W (2004) Improved MDS-based localization. In: *Proceedings of 23rd conference IEEE communications society (Infocom 2004)*, Hong Kong, 7–11 Mar 2004, pp 2640–2651
- Sun G, Chen J, Guo W, Ray Liu KJ (2005) Signal processing techniques in network aided positioning [A survey of state-of-the-art positioning designs]. *IEEE Signal Process Mag* 22:12–23
- Van MT, Tuan NV, Son TT, Le-Minh H, Burton A (2017) Weightedk-nearest neighbour model for indoor VLC positioning. *Commun IET* 11(6):864–871
- Wang J, Ghosh RK, Das SK (2010) A survey on sensor localization. *J Control Theory Appl* 8(1):2–11
- Wang B, Chen Q, Yang LT et al (2016) Indoor smartphone localization via fingerprint crowdsourcing: challenges and approaches. *IEEE Wirel Commun* 23(3):82–89
- Ward A, Jones A, Hopper A (1997) A new location technique for the active office. *IEEE Pers Commun Mag* 4(5):42–47
- Welch G, Bishop G, Vicci L, Brumback S, Keller K, Colucci D (1999) The Hi Ball tracker: high-performance wide-area tracking for virtual and augmented environments. In: *Symposium on virtual reality software and technology*, pp 1–10
- Wu KJ, Gregory TS, Moore J, Hooper B, Lewis D, Tse ZTH (2017) Development of an indoor guidance system for unmanned aerial vehicles with power industry applications. *Radar Sonar Navig IET* 11(1): 212–218
- Yan G, Olariu S (2009) An efficient geographic location-based security mechanism for vehicular adhoc networks. In: *2009 IEEE 6th international conference on mobile adhoc and sensor systems*, Macau, pp 804–809
- Ye A, Sao J, Jian Q (2016) A robust location fingerprint based on differential signal strength and dynamic linear interpolation. *Secur Commun Netw* 9:3618–3626
- Zadgaonkar H, Chandak M (2016) Object localization analysis using BLE: survey. *Smart Trends Inf Technol Comput Commun* 628:47
- Zou D, Meng W, Han S, Gong Z, Yu B (2013) User aided self-growing approach on radio map construction for WLAN based localization. In: *26th international technical meeting of the satellite division of the Institute of Navigation (ION GNSS + 2013)*, Nashville, Sept 2013

Location Based Services

- [Privacy Preservation for Location-Based Services](#)

Location Management

- [Mobility Management in 5G](#)

Longitudinal Data

- [Data Analytics for Longitudinal Biomedical Data](#)

Long-Term Evolution-Advanced

- [Smart Device-to-Device Communication: Communication Establishment and Maintenance](#)

Long-Term Quality of Protection

Wei Wang^{1,2}, Tao Jiang², and Qian Zhang³

¹ECE Department, University of Waterloo, Waterloo, ON, Canada

²Huazhong University of Science and Technology, Wuhan, China

³Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, People's Republic of China

Synonyms

Privacy preservation; Quality of protection

Definition

Long-term quality of protection (QoP) is envisioned to sustain personalized services with long-term privacy preservation.

Historical Background

Quality-of-protection (QoP) (Wang and Zhang 2015; Weber 2015) mechanisms have been exploited to thwart privacy breach in sharing user-generated content, including user's location information in location-based services (LBSs), and sensor data in participatory sensing (Wang et al. 2017). These mechanisms employ several QoP models that can be categorized into the following three classes (Wang and Zhang 2015).

Anonymization and generalization models. The crux of anonymization and generalization models is to replace a specific value of user's attributes (such as sensor data) with a coarser range to provide privacy protection (Zarrin et al. 2017). In LBS systems, the location coordinates of multiple users are transformed into the same geographical region to hide a user's specific location spot. In participatory sensing

systems, sensor data from multiple users can also be protected by releasing data of coarser granularity. The main advantage of anonymization and generalization models is that the generalized range truly includes the real value, and thus there is no falsified data. Whereas their limitations are also obvious: these techniques normally require a trusted third party to serve as the anonymizer. Moreover, these techniques work well under certain assumptions about adversaries' prior knowledge, while the users' privacy can be jeopardized in the presence of stronger adversaries, such as adversaries knowing the anonymization algorithm.

Randomization-based models. Randomization-based models include random perturbation and differential privacy techniques. Random perturbation transforms the original data by replacing a subset of the data points with randomly selected values, which can be generated by adding random noise to the original data or executing randomized algorithms. Differential privacy is achieved by injecting randomness into each user's data so that aggregated results are insensitive to the change of any single user's data. A major appealing feature of differential privacy is that it makes the worst case guarantee, that is, even if the adversaries know the data of all the individuals except the target individual, the adversaries are still uncertain about the data of the target individual. Though randomization-based models can defend a wider range of attacks than anonymization and generalization models, the loss of data quality and truthfulness caused by randomness is the main shortcoming.

Cryptographic techniques. End-to-end encryptions are widely adopted to secure the transmissions of users' reported data. These approaches mainly protect users' privacy from external adversaries (such as eavesdroppers), while internal adversaries (such as compromised or untrusted service providers) can decrypt the ciphertext to obtain users' data. Secure multiparty computation (SMC) allows multiple peers to jointly compute a function over their inputs without revealing inputs to each other.

However, SMC approaches are normally limited to specific computations and incur substantial communication or computation overhead.

Many studies leverage the above techniques to enable privacy preservation in LBSs or participatory sensing applications. Most of them are designed for protecting data privacy in single-shot scenarios, or focus on static attacks, while the user's two-level sensitivities and the context-aware adversaries cannot be covered by these techniques. In particular, the framework should be carefully designed to guarantee user's QoP requirements (context sensitivity and data sensitivity) against context-aware adversaries (passive and active adversaries) and different attacking strategies (static and adaptive attacks). To this end, QoP frameworks for long-term services are exploited to satisfy user's QoP requirements under different types of attacks.

The long-term QoP mechanisms are designed for long-term services where a user encounters a series of contexts and transits between various contexts (Wishart et al. 2015). The gamut of contexts is a set of interrelated user states and environmental conditions that are logically and temporally correlated (Gotz et al. 2012). The mobile user usually goes to a coffee shop after leaving home – this temporal correlation is the user's habit that can be learned from the user's historical data. The mobile user walks on the street after having coffee in the coffee shop – in this case, the context “walking” has logical relation with the context “in coffee shop” as a user needs to walk out of the coffee shop after having coffee.

Key Applications

Long-term QoP is designed to offer protection for long-term services. A range of sensors equipped on smart devices have been effectively used to infer user's contexts including location from GPS or wireless signals (Wi-Fi or cellular signals), mobility modes from accelerometer, and social activities from a combination of multiple sensors. Consequently, a large bundle of person-

alized services are emerging by utilizing user-generated sensor data. Smart devices carried by a mobile user continuously acquire readings from equipped sensors and periodically upload the data to corresponding service providers via the user's smartphone. After receiving the user-generated sensor data, service providers perform a series of actions to process the data, including context acquisition, preprocessing, storing, and reasoning within the application boundaries. The target of these actions is to “understand” the current state of the user and environmental conditions, based on which service providers deliver personalized services to the user. These personalized services can be categorized into:

Passive services. This type of services passively tracks the user's activities or conditions for offline analysis. Examples are healthcare monitoring and activity tracking applications such as Apple Healthbook and FitBit.

Active services. This type of services recognizes the user's contexts in real time and autonomously takes actions based on the current context. For example, an iOS application named AutoSilent automatically mutes the phone when the user is in a meeting.

Cross-References

► QoE Assessment Models

References

- Gotz M, Nath S, Gehrke J (2012) MaskIt: privately releasing user context streams for personalized mobile applications. *Proc ACM SIGMOD*:289–300
- Wang W, Zhang Q (2015) Toward long-term quality of protection in mobile networks: a context-aware perspective. *IEEE Wirel Commun* 22(4):34–40
- Wang S, Li L, Sun W, Guo J, Bie R, Lin K (2017) Context sensing system analysis for privacy preservation based on game theory. *Sensors* 17(2):339
- Weber RH (2015) Internet of things: privacy issues revisited. *Comput Law Secur Rev* 31(5):618–627
- Wishart R, Henricksen K, Indulska J (2007) Context privacy and obfuscation supported by dynamic context source discovery and processing in a context management system. In: *International conference*

on ubiquitous intelligence and computing. Springer, Berlin/Heidelberg, pp 929–940

Zarrin M, Houmansadr A, Pishro-Nik H (2017) Achieving perfect location privacy in wireless devices using anonymization. *IEEE Trans Inf Forensics Secur* 12(11):2683–2698

Low-Latency Communications with Millimeter Wave

Guang Yang¹ and Ming Xiao²

¹KTH Royal Institute of Technology, Stockholm, Sweden

²Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Low-latency millimeter-wave networks

Definitions

The overall latency in communication networks is defined as the time span of completing the transmission from the source to the destination, which generally comprises: propagation delay (sending one bit to receiver via the physical medium), transmission delay (pushing the packet into the communication medium in use), processing delay (analyzing a packet header, making a routing decision, and processing signals), and queuing delay (waiting in the queue for transmission).

Historical Background

Low latency is a critical feature in future wireless communications, i.e., the fifth-generation (5G) mobile networks, which can significantly improve the quality of service (QoS). It is

reported in Andrews et al. (2014) and Osseiran et al. (2014) that the anticipated 5G systems should enable the latency as low as a few milliseconds, and this requirement is more stringent than those in the existing 3G/4G systems, thereby posing huge challenges to related system development and implementation.

Fortunately, millimeter-wave (mm-wave) technologies as promising candidates in future mobile networks enable low-latency communications. The mm-wave radios in 5G networks usually refer to the electromagnetic waves with frequency among the spectra from 24 to 300 GHz. In mm-wave bands, the abundant spectral resources and dense antenna arrays (thanks to the short wavelength) can provide evidently higher data rates than those in sub-6 GHz bands, thereby paving a road for achieving low-latency communications. The features of mm-wave communications and the progress of mm-wave technologies are comprehensively summarized by Rappaport et al. (2013), Rangan et al. (2014), Xiao et al. (2017).

Driven by the ever-increasing demands on latency-sensitive applications and the huge benefits of mm-wave technologies, low-latency communications with mm-wave as a hot topic have attracted extensive attention. Numerous recent efforts have been devoted to the research on this topic, yielding many excellent achievements, e.g., Levanen et al. (2014), Filippou et al. (2017), Nielsen et al. (2018).

Related Works

In what follows, an overview that briefly summarizes several typical works in recent years regarding low-latency mm-wave communications is provided.

From the perspective of the transport-layer protocol, Polese et al. (2017) provided an overview regarding the impacts of mm-wave communications on the performance of the transmission control protocol (TCP), in terms of the end-to-end latency and reliability. Also, motivated by multiconnectivity in mm-wave

communications, the authors investigated the potential of the recent multipath TCP (MP-TCP) option in improving the network performance. It was revealed by Polese et al. (2017) that an optimized interaction between the transport-layer protocol and mm-wave network plays an important role in enabling low-latency 5G communications.

Using stochastic network calculus based on moment generating function (MGF), the probabilistic end-to-end latency was investigated by Yang et al. (2018a) for mm-wave multihop networks with full-duplex buffer-aided relays. Specifically, with lognormally distributed shadowing incorporated in mm-wave channels, upper bounds for the probabilistic end-to-end latency were developed to keep track of the latency performance, and a delay-optimum power allocation scheme subject to a sum-power constraint was also developed in the presence of self-interference. Yang et al. (2018a) provided an analytic method that can assess the probabilistic end-to-end latency of multihop networks in a straightforward manner, and also showed important clues for optimizing the power allocations in the presence of self-interference and constrained sum-power budget.

For the two distinct transmission schemes, namely, network densification (refer to serial transmission) and traffic dispersion (refer to parallel transmission), the low-latency capabilities were investigated by Yang et al. (2017) for mm-wave networks with buffers. In addition to the MGF-based network calculus method, effective capacity was adopted to characterize the service capability, which is related to the maximum traffic density that can be afforded by system without causing infinite latency. A hybrid scheme based on a flexible combination of network densification and traffic dispersion was also investigated. Given a fixed source-destination distance, it was shown by Yang et al. (2017) that the different low-latency capabilities of the three schemes heavily depend on the sum-power budget and the arrival rate.

Focusing on the massive multiple-input multiple-output (MIMO) network with mm-wave, Vu et al. (2017) investigated a trade-

off between capacity improvement and latency reduction. Specifically, the authors formulated a network utility maximization problem associated with probabilistic latency and reliability constraints and resorted to the Lyapunov method that adapts to channel variations and queue dynamics for seeking a utility-optimum solution. It was shown that the optimization problem can be decomposed into a dynamic latency control and rate allocation. Numerical results demonstrated that the approach proposed by Vu et al. (2017) exhibits promising gains in latency reduction.

Hashemi et al. (2018) proposed an architecture that integrates the sub-6 GHz and mm-wave technologies, exploiting the spatial correlations between interfaces for beamforming and data transfer, more precisely, the sub-6 GHz interface as a secondary option to assist the mm-wave interface. In addition, the authors formulated a scheduling problem regarding assigning the arrival traffic between the sub-6 GHz and mm-wave interfaces, where the two interfaces were treated with dedicated queues, as two individuals aiming to maximize the mm-wave throughput subject to a bounded delay. It was shown by Hashemi et al. (2018) that the optimal scheduling policy is threshold-based, which can be easily applied in spite of fast-varying mm-wave channels.

In the heterogeneous network (HetNet) with mm-wave, the latency performance of traffic allocation and cooperative networking was investigated by Yang et al. (2018b). Considering a simple HetNet that comprises one macro-cell evolved NodeB (MeNB), two small-cell evolved NodeBs (SeNBs), and one user equipment (UE), the authors resorted to the optimal traffic allocations and cooperation among MeNB and SeNBs, aiming to minimize the end-to-end latency in a downlink transmission. Yang et al. (2018b) revealed that traffic allocation and cooperative networking are capable of significantly reducing the latency in mm-wave HetNets, while the complexity will be a huge concern as the network size scales up.

In addition to the above studies towards specific technical problems, there are several

surveying works summarizing the state-of-the-art research regarding low-latency communications in 5G from various aspects. Recently, a survey on the emerging technologies for low latency towards 5G was presented by Parvez et al. (2017), where the potential solutions in the three domains, i.e., radio access network (RAN), core network, and caching, were extensively investigated. Another comprehensive survey was presented by Ford et al. (2017), which summarized some challenges and potential solutions for providing reliable and low-latency services in mm-wave systems. The state-of-the-art research was reviewed from the following three aspects in high-layer designs: core network architecture, medium access control (MAC), and congestion control. It was concluded by Ford et al. (2017) that, to achieve low latency in mm-wave communications, it is crucial to make data and services physically closer to the end user and to enable flexible and scalable network deployment and functionality management. In addition, flexible MAC layer with low-latency scheduling and fast adaptive congestion control against fast-varying mm-wave channels are important.

Key Applications

Low latency plays an extremely important role in a broad class of use cases, especially for those latency-sensitive and/or mission-critical applications, including, e.g., virtual reality (VR), augmented reality (AR), intelligent transportation systems (ITS), tactile Internet, vehicle-to-vehicle (V2V), industrial control and automation, remote healthcare applications, robotics, and telepresence.

Cross-References

- ▶ [Millimeter-wave Communications](#)
- ▶ [Quality-of-Service \(QoS\)](#)
- ▶ [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

References

- Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong AC, Zhang JC (2014) What will 5G be? *IEEE J Sel Areas Commun* 32(6):1065–1082
- Filippou MC, De Donno D, Priale C, Palacios J, Giustiniano D, Widmer J (2017) Throughput vs. latency: QoS-centric resource allocation for multi-user millimeter wave systems. In: *IEEE international conference on communications (ICC)*, IEEE, pp 1–6
- Ford R, Zhang M, Mezzavilla M, Dutta S, Rangan S, Zorzi M (2017) Achieving ultra-low latency in 5G millimeter wave cellular networks. *IEEE Commun Mag* 55(3):196–203
- Hashemi M, Koksall CE, Shroff NB (2018) Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems. *IEEE Trans Commun* 66(2): 875–888
- Levanen T, Pirskanen J, Valkama M (2014) Radio interface design for ultra-low latency millimeter-wave communications in 5G era. In: *IEEE globecom workshops (GC Wkshps)*, IEEE, pp 1420–1426
- Nielsen JJ, Liu R, Popovski P (2018) Ultra-reliable low latency communication using interface diversity. *IEEE Trans Commun* 66(3):1322–1334
- Osseiran A, Boccardi F, Braun V, Kusume K, Marsch P, Maternia M, Queseth O, Schellmann M, Schotten H, Taoka H, Tullberg H, Uusitalo MA, Timus B, Fallgren M (2014) Scenarios for 5G mobile and wireless communications: the vision of the METIS project. *IEEE Commun Mag* 52(5):26–35
- Parvez I, Rahmati A, Guvenc I, Sarwat AI, Dai H (2017) A survey on low latency towards 5G: RAN, core network and caching solutions. *CoRR*. arXiv preprint arXiv:170802562
- Polese M, Jana R, Zorzi M (2017) TCP and MP-TCP in 5G mmWave networks. *IEEE Internet Comput* 21(5):12–19
- Rangan S, Rappaport TS, Erkip E (2014) Millimeter-wave cellular wireless networks: potentials and challenges. *Proc IEEE* 102(3):366–385
- Rappaport TS, Sun S, Mayzus R, Zhao H, Azar Y, Wang K, Wong GN, Schulz JK, Samimi M, Gutierrez F (2013) Millimeter wave mobile communications for 5G cellular: it will work! *IEEE Access* 1: 335–349
- Vu TK, Liu CF, Bennis M, Debbah M, Latva-aho M, Hong CS (2017) Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks. *IEEE Commun Lett* 21(9):2041–2044
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis GK, Bjornson E, Yang K, Chih-Lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun* 35(9):1909–1935
- Yang G, Xiao M, Poor HV (2017) Low-latency millimeter-wave communications: traffic dispersion

or network densification? CoRR. arXiv preprint arXiv:170908410

Yang G, Xiao M, Al-Zubaidy H, Huang Y, Gross J (2018a) Analysis of millimeter-wave multi-hop networks with full-duplex buffered relays. *IEEE/ACM Trans Networking* 26(1):576–590

Yang G, Xiao M, Alam M, Huang Y (2018b) Low-latency heterogeneous networks with millimeter-wave communications. CoRR. arXiv preprint arXiv:180109286

Low-Latency Millimeter-Wave Networks

- [Low-Latency Communications with Millimeter Wave](#)

Low-Latency Smart Grid Communications

- [Ultra-reliable and Low-Latency Communications for the Smart Grid](#)

LTE: Long-Term Evolution

- [Security and Privacy in 4G/LTE Network](#)

LTE-Unlicensed (LTE-U)

- [Effective Utilization of Unlicensed Spectrum](#)

MAC in Cognitive Radio Networks

Juncheng Jia
Soochow University, Suzhou, China

Synonyms

[Media access control protocol for cognitive radio networks](#)

Definitions

In cognitive radio networks, the media access control (MAC) protocols need to determine the availability of spectrum resource through spectrum sensing and coordinate with the other secondary users (SUs) for spectrum access, so that the spectrum resource is efficiently utilized without harmful interference to the primary users (PUs).

Historical Background

Cognitive radio has been first proposed in Mitola and Maguire (1999) and Mitola (2000) as a way to promote the efficient use of the spectrum by exploiting spectrum opportunities. SUs with cognitive radios are able to exploit information about the wireless environment and adapt their

transmission parameters to access available channels while avoiding harmful interference to PUs (Haykin 2005).

In general, there are three transmission schemes for cognitive radios: underlay, overlay, and interweave (Goldsmith et al. 2009). In underlay scheme, SUs are allowed to transmit signal as long as the generated interference stays below a certain threshold. In 2003 the FCC defined the interference temperature as a way to measure and limit the interference at PUs. However, this model has been abandoned by the FCC in 2007 due to the implementation issues. In overlay scheme, SUs exploit the knowledge of PUs' messages to either limit the interference. In interweave scheme, SUs can only transmit in spectrum holes; whenever a SU detects PUs, the SU vacates its channel to avoid harmful interference to PUs.

In order to realize the functions of cognitive radios, MAC should possess the ability of channel sensing, resource allocation, spectrum sharing, and spectrum mobility (Akyildiz et al. 2006). Channel sensing is the ability of a cognitive radio to collect information about spectrum usage. Resource allocation is employed to opportunistically assign available channels to SUs according to QoS requirements. Spectrum sharing deals with contentions between heterogeneous PUs and SUs in order to avoid harmful interference. Spectrum mobility allows a SU to vacate its channel, when a PU is detected, and to access an idle band where it can reestablish the communication link. The research works

of cognitive radio MAC mainly focus on these aspects, which is reviewed in several surveys (Krishna and Amitabha 2009; De Domenico et al. 2012; Cormio and Chowdhury 2012). Apart from the close coupling of physical layer and MAC layer, there also exist interactions between the network and transport layers with MAC layer, such as joint spectrum and route decisions, congestion-free end-to-end reliability, spectrum and node mobility, etc. (Akyildiz et al. 2009).

On the standardization side of cognitive radio, IEEE 802.22 represents the major effort, which was started in 2004 and finalized in 2011 (Cordeiro et al. 2005). 802.22 is designed for wireless regional area networks (WRANs) which takes advantage of the favorable transmission characteristics of the VHF and UHF TV bands to provide broadband wireless access over a large area up to 100 km from the transmitter. 802.22 MAC uses time division multiplexing, with structures of superframe and frame, where a superframe is composed of multiple MAC frames preceded by the frame preamble. In addition to 802.22, the IEEE has standardized another white space cognitive radio standard, 802.11af, which was approved in 2014 (Flores et al. 2013). 802.11af is a wireless LAN standard designed for ranges up to 1 km. 802.11af uses carrier sense multiple access with collision Avoidance (CSMA/CA)-based MAC protocol.

Foundations

Due to the unique requirement of protection for PUs, the design of cognitive radio MAC protocols is different from MAC protocols of most of the other wireless systems. The cognitive radio MACs are based on the closed coupling with the physical layer and the hardware support on the radio device.

Spectrum sensing is one of the key enabling functions in cognitive radio networks that is used to explore vacant spectrum opportunities and to avoid interference with the PUs. While the physical layer of spectrum sensing is responsible for the signal processing work, the MAC layer

is more involved in the higher-level functions of spectrum sensing coordination. For example, MAC layer needs to optimize sensing and transmission durations, the order in which channels are sensed, and the number of channels to be sensed before transmission (Kim and Shin 2008; Lee and Akyildiz 2008; Jia et al. 2008). Furthermore, the reliability of spectrum sensing can be improved if SUs share their sensing results with neighbors due to multiuser spatial diversity, which is called cooperative spectrum sensing. However, the improvement of spectrum sensing comes at the cost of increased latency and communication overhead, where the MAC layer tries to coordinate multiple SUs and reduce the cost.

Since cognitive radio networks usually use multiple channels, control channel establishment needs to be carefully considered, especially for distributed networks. Some MAC designs assume there exists a global control channel available for all SUs, which simplifies the implementation (Ma et al. 2005). When a global control channel is not available or reliable, local control channels can be used, since it is possible that neighboring SUs may share some commonly available channels (Zhao et al. 2005; Chen et al. 2005). Besides, some designs do not use a dedicated control channel at all, which is based on channel hopping. The coordination of channel hopping in such context is usually referred to as rendezvous (Lin et al. 2011; Bian et al. 2011).

Spectrum allocation is an important function of the MAC protocol in cognitive radio networks. Since the spectrum availability and transmission requirement at SUs could be different, SUs have to share their spectrum availability information with neighbors and conduct efficient spectrum allocation to maximize certain utility. MAC protocols exploit advanced optimization algorithms to realize intelligent, fair, and efficient allocation of the available spectrum. Each SU adapts its transmission parameters to changes of the wireless environment, in order to efficiently exploit the available resource. Considering the high complexity of centralized optimization, decentralized approaches in which each SUs acts based on partial knowledge of network status have been

proposed based on graph coloring theory (Zheng and Peng 2005), game theory (Wang et al. 2010), stochastic theory (Zhao et al. 2007), etc.

Spectrum access enables multiple SUs to share the spectrum resource by determining who will access the channel or when a user accesses the channel. Contentions between SUs can be avoided through coordinated access both in centralized and distributed architectures. When coordination is absent, a random approach could be exploited to contend for access to available channels (Zhao et al. 2007). Otherwise, SUs may exchange signalling messages in order to reserve the access to a data channel (Su and Zhang 2008). The control handshaking mechanism, however, does not completely solve the hidden terminal problem; hence the busy tone scheme is often exploited to prevent hidden nodes (Ma et al. 2005).

When a PU is detected, to realize seamless transmission, a cognitive radio vacates its channel and reconstructs a transmission link on a different channel. The procedure that permits this transition from a channel to another with minimum performance degradation is called handoff. Spectrum mobility in MAC layer is to reduce delay and loss during spectrum handoff. In general, there are two different strategies for spectrum mobility: proactive spectrum handoff and reactive spectrum handoff (Akyildiz et al. 2009). In proactive spectrum handoff, SUs predict future activity in the current link and determine a new spectrum while maintaining the current transmission and then perform spectrum switching before the link failure happens. For proactive spectrum handoff, the spectrum switching is faster but requires complex algorithms. On the other hand, in reactive spectrum handoff, SUs perform spectrum switching after detecting link failure due to spectrum mobility. This method requires immediate spectrum switching without any preparation time, resulting in significant quality degradation in ongoing transmissions. It is also possible to design a hybrid spectrum handoff strategy, which combines proactive strategy and reactive strategy by applying proactive spectrum sensing and reactive handoff action (Christian et al. 2012).

Key Applications

MAC protocol is a fundamental component for cognitive radio networks.

Cross-References

- [Cognitive Heterogeneous Networks](#)
- [Millimeter Wave MAC Layer](#)
- [QoS-Aware MAC](#)

References

- Akyildiz IF, Lee WY, Vuran MC, Mohanty S (2006) NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey. *Comput Netw J* 50:2127–2159
- Akyildiz I, Lee W, Chowdhury K (2009) CRAHNs: cognitive radio ad hoc networks. *Ad Hoc Netw* 7:810–836
- Bian K, Park J, Chen R (2011) Control channel establishment in cognitive radio networks using channel hopping. *IEEE J Sel Areas Commun* 29:689–703
- Chen T, Zhang H, Maggio G, Chlamtac M (2005) CogMesh: a cluster-based cognitive radio network. In: *IEEE DySPAN*, pp 168–178
- Christian I, Moh S, Chung I, Lee J (2012) Spectrum mobility in cognitive radio networks. *IEEE Commun Mag* 50:114–121
- Cordeiro C, Challapali K, Birru D, Shankar S (2005) IEEE 802.22: the first worldwide wireless standard based on cognitive radios. In: *IEEE DySPAN*, pp 328–337
- Cormio C, Chowdhury K (2012) A survey on MAC protocols for cognitive radio networks. *Ad Hoc Netw* 7:1315–1329
- De Domenico A, Strinati EC, Di Benedetto MG (2012) A survey on MAC strategies for cognitive radio networks. *IEEE Commun Surv Tutor* 14:21–44
- Flores A, Guerra R, Knightly E, Ecclesine P, Pandey S (2013) IEEE 802.11 af: a standard for TV white space spectrum sharing. *IEEE Commun Mag* 51:92–100
- Goldsmith A, Jafar SA, Maric I, Srinivasa S (2009) Breaking spectrum gridlock with cognitive radios: an information theoretic perspective. *Proc IEEE* 97:894–914
- Haykin S (2005) Cognitive radio: brain-empowered wireless communications. *IEEE J Sel Areas Commun* 23:201–220
- Jia J, Zhang Q, Shen X (2008) HC-MAC: a hardware constrained cognitive MAC for efficient spectrum management. *IEEE J Sel Areas Commun* 26:106–117
- Kim H, Shin K (2008) Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks. *IEEE Trans Mob Comput* 7:533–545

- Krishna TV, Amitabha D (2009) A survey on MAC protocols in OSA networks. *Comput Netw* 9:1377–1394
- Lee W, Akyildiz I (2008) Optimal spectrum sensing framework for cognitive radio networks. *IEEE Trans Wirel Commun* 7:845–857
- Lin Z, Liu H, Chu X, Leung Y (2011) Jump-stay based channel-hopping algorithm with guaranteed rendezvous for cognitive radio networks. In: *IEEE INFOCOM*, pp 2444–2452
- Ma L, Han X, Shen CC (2005) Dynamic open spectrum sharing MAC protocol for wireless ad hoc networks. In: *IEEE DySPAN*, pp 203–213
- Mitola J (2000) Cognitive radio: an integrated agent architecture for software defined radio. PhD thesis, Royal Institute of Technology (KTH), Stockholm
- Mitola J, Maguire G (1999) Cognitive radio: making software radios more personal. *IEEE Pers Commun* 6:13–18
- Su H, Zhang X (2008) Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks. *IEEE J Sel Areas Commun* 26:118–129
- Wang B, Wu Y, Liu K (2010) Game theory for cognitive radio networks: an overview. *Comput Netw* 54:2537–2561
- Zhao J, Zheng H, Yang GH (2005) Distributed coordination in dynamic spectrum allocation networks. In: *IEEE DySPAN*, pp 259–268
- Zhao Q, Tong L, Swami A, Chen Y (2007) Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework. *IEEE J Sel Areas Commun* 25:589–600
- Zheng H, Peng C (2005) Collaboration and fairness in opportunistic spectrum access. In: *IEEE ICC*, pp 3132–3136

Machine Learning

- [Application of Machine Learning in Wireless Sensor Network](#)
- [Big Data in 5G](#)
- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)
- [Machine Learning in Wireless Sensor Networks for the Internet of Things](#)

Machine Learning Algorithms

- [Machine Learning Paradigms in Wireless Network Association](#)

Machine Learning in Wireless Sensor Networks for the Internet of Things

Abdallah Jarwan¹, Ayman Sabbah²,
and Mohamed Ibnkahla³

¹Carleton University, Ottawa, ON, Canada

²Spectrum and Telecommunication Sector (STS) - Innovation, Science, and Economic Development (ISED)/Government of Canada, Ottawa, ON, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

[Artificial intelligence](#); [Machine learning](#)

Definition

In usual ways of programming, a program is built by building instructions in order to reach desired outputs from inputs. However, in machine learning (ML), that process is flipped. The inputs and desired set of outputs are given, and the program should learn what instructions or policy should be followed. ML is concerned with algorithms that observe data, learn from it, grow up, and make more intelligent decisions. Based on the way that machines accumulate knowledge and become able to function as needed, ML can be classified into supervised, semi-supervised, unsupervised, and reinforcement learning.

Historical Background

Internet of Things (IoT) systems are built on thousands of wireless sensor networks (WSNs). WSNs suffer from various limitations such as finite energy, hardware inaccessibility, uncontrolled environment, uncontrolled topology, limited computational power and storage capability,

and heterogeneity in systems. Recent developments of WSN technologies have targeted energy efficiency, lifetime, coverage, connectivity, traffic balancing, spectrum management, fault tolerance, latency, reliability, and security, among others. These targets can be improved by considering controlled deployment and mobility, routing, scheduling, power and rate control, clustering, data aggregation, in-network processing, and traffic control. One of the most important aspects of WSNs and IoT systems is traffic control through traffic reduction. Traffic reduction is done by reducing the amount of data or packets that need to be transferred across WSNs and various IoT layers. This improves IoT systems in terms of energy and spectrum, which are the main resources, and increases lifetime. Traffic control can be implemented in IoT systems using machine learning (ML) algorithms.

ML algorithms observe data, learn from it, grow up, and make intelligent decisions. Through training, machines build policies to perform complex tasks that mimic intelligent human behavior. ML tools have the potential to provide lower-complexity alternatives for sophisticated algorithms that consume time and resources. Also, ML makes it easy to deduce hidden correlations, understand context, and make predictions from the collected data. It also provides autonomous and intelligent decision making schemes with minimal human intervention. The importance of applying ML approaches in IoT systems and WSNs is mainly due to the dynamicity and uncontrollability of the surrounding environment (Alsheikh et al. 2014). Deep neural networks (DNNs) and long short-term memory networks (LSTMs) are supervised ML techniques. They are used to improve WSNs in various applications. In this chapter, both will be studied and proposed as a key application for traffic reduction in IoT systems.

The remainder of this chapter is organized as follows. First, mathematical computational graphs of DNNs and LSTMs are used to illustrate how they work. After that, their applications to WSNs and IoT systems are studied. Finally, both techniques are used to improve traffic reduction through IoT systems.

Foundations

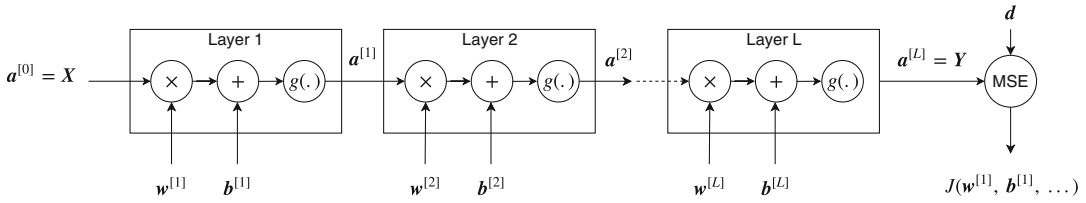
Deep Neural Networks (DNNs)

Deep learning is a machine learning technique where learning is done through multiple levels of representations (Chen and Lin 2014). Deep neural networks (DNNs) mimic how humans' brains work in the sense that they are composed of many abstraction layers. Each layer draws conclusions based on the previous layer outputs. They are widely used in feature extraction, modeling complex relationships between inputs and outputs, and context understanding.

DNNs are represented in computational graphs as shown in Fig. 1. The illustrated DNN consists of L layers. The l th layer takes an $n^{[l-1]} \times 1$ vector $\mathbf{a}^{[l-1]}$ and outputs another $n^{[l]} \times 1$ vector $\mathbf{a}^{[l]}$ that goes as an input to the next layer, where $n^{[l]}$ is the number of neurons at the l th layer. The output of each layer is computed from its input. In the figure, $\mathbf{w}^{[l]}$ is the l th layer $n^{[l]} \times n^{[l-1]}$ weight matrix, $\mathbf{b}^{[l]}$ is its $n^{[l]} \times 1$ bias, and $g(\cdot)$ is an element-wise activation function. The activation function $g(\cdot)$ can be Sigmoid, tanh, rectified linear unit (ReLU) functions, among others.

As a supervised learning approach, the DNN is trained using a labeled data set. The goal of building a DNN is to map a new input $\mathbf{X}$ to an output $\mathbf{Y}$ based on the labeled inputs. During this phase, a cost function $J(\mathbf{w}^{[1]}, \mathbf{b}^{[1]}, \mathbf{w}^{[2]}, \mathbf{b}^{[2]}, \dots)$ is defined in terms of outputs $\mathbf{Y}$ and labels $\mathbf{d}$ as shown in Fig. 1. For example, the cost can be defined as the mean square error (MSE).

The DNN output $\mathbf{Y}$ is computed using the computational graph in Fig. 1. This process is called forward propagation. During training, the gradients $d\mathbf{w}^{[l]} = \partial J / \partial \mathbf{w}^{[l]}$ and $d\mathbf{b}^{[l]} = \partial J / \partial \mathbf{b}^{[l]}$ at every layer are computed. After that, they are used to update (tune) the weights and biases in a process that is called backward propagation. Weights and biases updates are done using various iterative optimization algorithms such as gradient descent (GD) or Adam optimizer. In GD algorithm, the updates are done using $\mathbf{w}^{[l]} := \mathbf{w}^{[l]} - \alpha d\mathbf{w}^{[l]}$ and $\mathbf{b}^{[l]} := \mathbf{b}^{[l]} - \alpha d\mathbf{b}^{[l]}$, where $:=$ denotes the update operator and α represents the learning rate. Gradients and updates are done



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 1 DNNs computational graph architecture

numerically using frameworks such as TensorFlow (Rampasek and Goldenberg 2016). It is emphasized that TensorFlow is optimized to calculate gradients because it is based on computational graphs where the chain rule is used to compute derivatives.

Long Short-Term Memory Networks (LSTMs)

Long Short-Term Memory networks (LSTMs) represent a special kind of recurrent neural networks (RNNs). LSTMs are good in handling sequence dependencies where the inputs are represented in terms of time, such as in prediction and language processing problems. An LSTM network is composed of a single LSTM cell that has a feedback connection from its output to its input. The structure of an LSTM cell is shown in Fig. 2. It is shown that at any time instance t , the LSTM cell input is composed of the current input (x_t) and the outputs from the previous time instance (h_{t-1} and c_{t-1}). LSTMs can be mathematically considered as multiple cells that are connected in cascade as shown in Fig. 3. However, it differs from DNNs in the sense that all LSTM cells in the cascade connection use the same weights and biases.

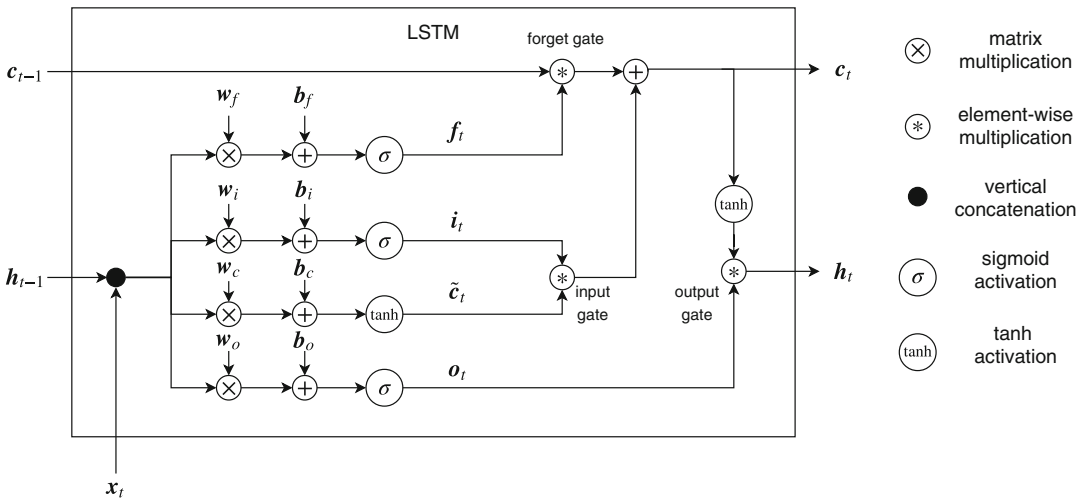
The LSTM cell contains four NNs that are connected to form three gates: forget (f), input (i), and output (o) gates as shown in Fig. 2. Each NN comprises one neurons layer with H number of neurons. The weights and biases of these NNs are denoted by w_f , b_f , w_i , b_i , w_c , b_c , w_o , and b_o . In general, during the training phase, these variables are updated so that the LSTM performs as required. Each LSTM cell has a state c_t and an output h_t at time instance t . Both cell state and output are fed to the LSTM in the next time

instance $t + 1$. The new cell state and output are calculated from previous state, previous output, and current input using the computational graph in Fig. 2. The cost function and variable tuning (backward propagation) can be done similar to what has been illustrated in DNNs.

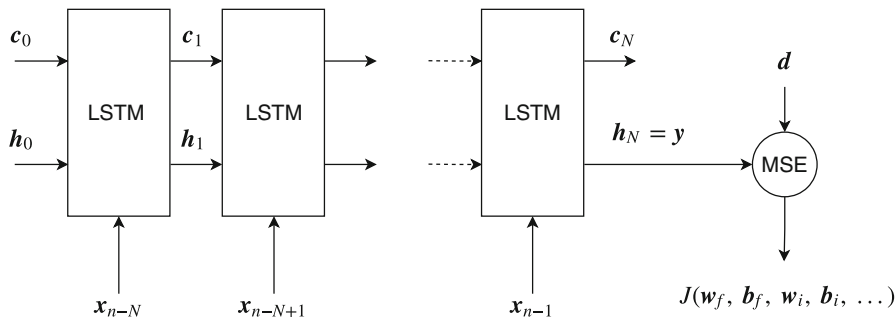
Key Applications

Applications of DNNs and LSTMs in IoT Systems

DNNs are used to tackle many applications in IoT (Chen et al. 2017). For instance, DNNs are powerful for big data analytics (Mohammadi et al. 2018). Therefore, they are used to extract patterns and conclusions from data gathered by IoT devices. For example, DNNs can be used in context awareness applications such as human activity recognition in wearable WSNs. Moreover, DNNs are used to make predictions using simple information as input. Also, they can be used for classification problems which can be applied to many aspects in WSNs and IoT. Furthermore, DNNs have many applications in automation and building self-organizing networks which are the essence of IoT systems. The recent literature has proposed many applications of DNNs in applications related to IoT systems. In Bande and Shete (2017), DNNs are used for flood management in IoT systems to predict floods. The proposed framework has been applied to data collected from Chennai region. In Kumar and Jain (2018), localization in IoT scenarios has been implemented using DNNs. Localization is critical to many application where nodes location should be monitored, while GPS devices are not feasible to be used. DNNs are



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 2 Computational graph of a single LSTM cell



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 3 Prediction architecture using LSTMs

used to improve tracking in WSNs as well, as discussed in Luo et al. (2016).

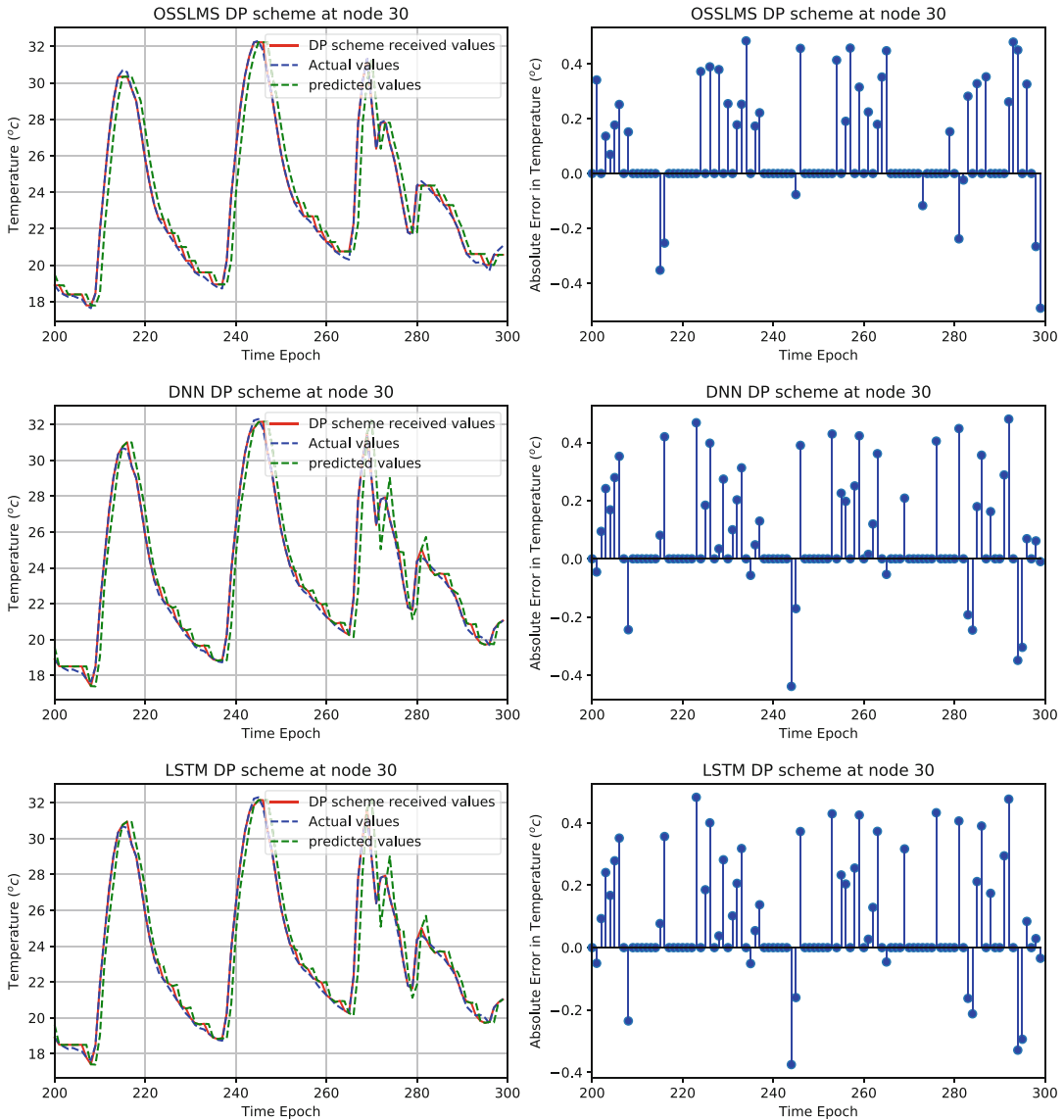
Due to their gated and memory architecture, LSTMs are suitable for extracting patterns where the input data spans over long sequences. Therefore, LSTMs are suitable for prediction and time-series analytics (Jansen et al. 2018). They can be also used to extract features from long sequences (Liu et al. 2016). Moreover, LSTMs have some implementations for filling missing data points in data sets (Lipton et al. 2016). In literature, many LSTM applications can be found in WSNs and IoT research. For instance, LSTMs are commonly used architectures for short-term traffic flow prediction as illustrated in Ali and Mahmood (2018). In Aydin and Guldamlasioglu (2017), LSTMs

are used for maintenance systems in predicting the condition of components in factories and autonomous IoT systems.

In order to illustrate the applicability of using DNNs and LSTMs in IoT systems, they are applied to traffic reduction problem. Next, the dual prediction scheme and its performance when applying DNNs and LSTMs will be studied.

Traffic Reduction in IoT Systems Using Dual Prediction

Most of the collected data is usually redundant and can be mined from other observations due to the existence of a temporal correlation. Preventing the unnecessary transmissions in a WSN has a significant impact on reducing

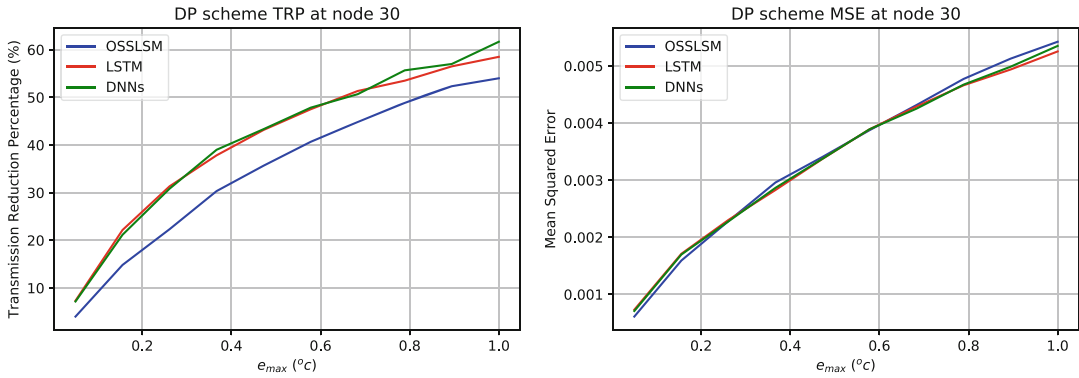


Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 4 DP scheme at Node 30 using OSSLMS, DNNs, and LSTMs

energy consumption and bandwidth usage. One approach to reduce data transmission between any two endpoints is to use dual prediction (DP) (Wu et al. 2016). The endpoints can be a node-to-node, node-to-gateway, or node-to-cloud. The DP scheme is based on data prediction algorithms that can be implemented using DNNs or LSTMs.

The two ends of the DP scheme are a source of data (S) and a destination (D). Let us assume

that a data buffer of size N is used to hold the last N observations at S and D. The DP scheme at S can be explained as follows. At the n th time slot, the data buffer is represented as $X_n = [x_{n-1}, x_{n-2}, \dots, x_{n-N}]$ where x_{n-i} represents the data point in the buffer at the $(n-i)$ th time slot. Initially, for the first N time slots, collected observations are transmitted from S to D and placed in the data buffer X_n . After that, when observation O_n is made at time slot n , the



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 5 DP scheme performance while using different accuracy levels

information in X_n is used to predict the observed value. The implemented prediction algorithms take X_n as input and generates a prediction P_n at time slot n . If the predicted value P_n was not close enough to the observed value O_n ($|P_n - O_n| > e_{\max}$ where $e_{\max}$ is the maximum acceptable prediction error), then the data buffer will be updated as $X_{n+1} = [O_n, x_{n-1}, \dots, x_{n-N+1}]$. Also, O_n is transmitted to D, and the value O_n is used to update the prediction model variables. However, if the predicted value P_n was close enough to the observed value O_n ($|P_n - O_n| \leq e_{\max}$), then $X_{n+1} = [P_n, x_{n-1}, \dots, x_{n-N+1}]$ and no transmission occurs because this observation can be predicted accurately. Also, the value P_n is used to update the prediction model. Following this scheme at S eliminates the need for transmitting the observations that can be accurately predicted at D.

The DP scheme at D is implemented as follows. During the first N time slots, the received observations are just saved at buffer X_n . After that, whenever an observation O_n is received, it is saved in the buffer as the newest data point and used to update the prediction model. However, if no observation is received at time slot n , then it means that this point can be predicted accurately; therefore, the prediction model is used to generate the corresponding observation. After that, the predicted point is used to update the prediction model and is added to the data buffer. It should be noted that the transmitting and receiving ends use the same prediction model. Also, they perform

the same model updates; therefore, both models are synchronized.

To evaluate our proposed models, the results are compared with optimal step-size least mean square (OSSLMS) algorithm that is discussed in Wu et al. (2016). The simulation has been performed on sample data collected by the Intel Berkeley research lab during a 1-month period that includes temperature, humidity, and light intensity. It is assumed that one node in Intel's WSN tries to send temperature readings periodically to another end which can be a cluster head, edge node, or cloud. It is assumed to be the cloud here. In any case, the DP scheme has the same performance.

The following parameters have been set for evaluation: the acceptable error threshold is $e_{\max} = 0.5^\circ\text{C}$ and the data buffer size $N = 10$. Figure 4 shows the performance of the DP scheme when using OSSLMS, DNNs, and LSTMs for predictions. It shows how the observed (at Node 30), predicted (at Node 30 and the cloud), and perceived (at the cloud) data changes over 100 time epochs. It can be seen that DP scheme's perceived data points (red) are equal to the predicted values (green) when the prediction is accurate and equal to the observed values (blue) when the prediction is not. Also, the error between the observed values (at Node 30) and the perceived ones (at the cloud) does not exceed $e_{\max}$ at any point. Figure 5 shows the percentage of traffic reduction and the MSE while varying the acceptable

error value $e_{\max}$. It shows how the proposed DNN and LSTM algorithms outperform recent literature algorithms such as OSSLMS. While all have the same MSE performance, DNN and LSTM outperform OSSLMS in terms of traffic reduction.

Cross-References

- [Application of Machine Learning in Wireless Sensor Network](#)
- [Machine Learning in Wireless Sensor Networks for the Internet of Things](#)

References

- Ali U, Mahmood T (2018) Using deep learning to predict short term traffic flow: a systematic literature review. In: Intelligent transport systems – from research and development to the market uptake. Springer International Publishing, Cham, pp 90–101
- Alsheikh MA, Lin S, Niyato D, Tan HP (2014) Machine learning in wireless sensor networks: algorithms, strategies, and applications. *IEEE Commun Surv Tutor* 16(4):1996–2018
- Aydin O, Guldamlasioglu S (2017) Using LSTM networks to predict engine condition on large scale data processing framework. In: 2017 4th international conference on electrical and electronic engineering (ICEEE), pp 281–285
- Bande S, Shete V (2017) Smart flood disaster prediction system using IoT neural networks. In: International conference on smart technologies for smart nation (SmartTechCon), pp 189–194
- Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. *IEEE Acc* 2:514–525
- Chen M, Challita U, Saad W, Yin C, Debbah M (2017) Machine learning for wireless networks with artificial intelligence: a tutorial on neural networks. *arXiv preprint:171002913*
- Jansen F, Holenderski M, Ozcelebi T, Dam P, Tijsma B (2018) Predicting machine failures from industrial time series data. In: 5th international conference on control, decision and information technologies (CoDIT), pp 1091–1096
- Kumar A, Jain V (2018) Feed forward neural network-based sensor node localization in internet of things. In: Pattnaik PK et al (eds) *Progress in computing, analytics and networking*. Springer, Singapore, pp 795–804
- Lipton ZC, Kale D, Wetzel R (2016) Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In: *Proceedings of Machine learning for healthcare conference*, PMLR, vol 56, pp 253–270
- Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal LSTM with trust gates for 3D human action recognition. In: *European conference on computer vision*. Springer, Amsterdam, pp 816–833
- Luo X, Lv Y, Zhou M, Wang W, Zhao W (2016) A laguerre neural network-based ADP learning scheme with its application to tracking control in the internet of things. *Pers Ubiquit Comput* 20(3):361–372
- Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M (2018) Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys Tutorials Early Access*
- Rampasek L, Goldenberg A (2016) Tensorflow: biology’s gateway to deep learning? *Cell Syst* 2(1):12–14
- Wu M, Tan L, Xiong N (2016) Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. *Info Sci* 329:800–818, special issue on Discovery Science

Machine Learning Paradigms in Wireless Network Association

Jingjing Wang¹ and Chunxiao Jiang²

¹Department of Electronic Engineering, Tsinghua University, Beijing, People’s Republic of China

²Department of Electronic Engineering, Tsinghua Space Center, Tsinghua University, Beijing, People’s Republic of China

Synonyms

[Cooperative wireless networks](#); [Machine learning algorithms](#); [Resource allocation](#)

Definitions

Investigating machine learning aided cooperative resource allocation and network association mechanisms for wireless networks.

Introduction

Wireless network association has received substantial attention both in the academic

and industrial communities. One of their driving forces is that of efficiently providing unprecedented data rates for supporting radical new applications (Jiang et al. 2017). Specifically, a range of network association schemes are expected to learn the diverse and colorful characteristics of both the users' ambience and the human behaviors, in order to autonomously determine the optimal system configurations for the sake of conserving energy as well as maximizing network capacity. Moreover, smart mobile terminals have to rely on sophisticated learning and decision-making. Machine learning, as one of the most powerful artificial intelligence tools, constitutes a promising solution for wireless network association (Alsheikh et al. 2014).

Machine learning has found wide-ranging applications in image/audio processing, finance and economics, social behavior analysis, etc. Explicitly, a machine learns the execution of a particular task **T**, with the goal of maintaining a specific performance metric **P**, based on a particular experience **E**, where the system aims for reliably improving its performance **P** while executing task **T**, again by exploiting its experience **E**. Machine learning algorithms can be simply categorized as *supervised learning* and *unsupervised learning*, where the adjectives "supervised/unsupervised" indicate whether there are labeled samples in the database. Later, *reinforcement learning* emerged as a new category, which was inspired by behavioral psychology. It is concerned with an agent's certain form of reward/utility, who is connected to its environment via perception and action. Relying on a cascade of multiple layers of nonlinear processing units for feature extraction and transformation, *deep learning* was applied to network association for the sake of its expressive capacity and convenient optimization capability compared with conventional machine learning algorithms.

In wireless networks, machine learning can be widely used in modeling various technical problems of large-scale MIMOs, device-to-device (D2D) networks, heterogeneous networks (Het-Nets), cognitive radio, etc. (Clancy et al. 2007).

In this entry, we will introduce the basic concept of machine learning algorithms and their corresponding applications in network association according to the category of supervised, unsupervised, reinforcement learning, and deep learning.

Supervised Learning in Network Association

Models

The *k*-nearest neighbor (KNN) and *support vector machines* (SVM) are conceived for classification of points/objects relying on a D-dimensional vector **x** of input variables. In KNN, an object is classified into a specific category by a majority vote of the object's neighbors, with the object being assigned to the class that is most common among its *k*-nearest neighbors. By contrast, the SVM relies on a nonlinear mapping, which transforms the original training data into a higher dimension where it becomes separable and then it searches for the optimal linear separating hyperplane that is capable of separating one class from another in this higher dimension.

The philosophy of *Bayesian learning* is to compute the a posteriori probability distribution of the target variables conditioned on its input signals and on all of the training instances. Some simple examples of generative models that may be learned with the aid of Bayesian techniques include, but are not limited to, the *Gaussian mixture model* (GM), *expectation maximization* (EM), and *hidden Markov models* (HMM). Specifically, GM is a model where each data point belongs to one of several clusters or groups, and the data points within each cluster are Gaussian distributed. EM is a generalization of maximum likelihood estimation, which iteratively finds the most likely solutions or parameters. It is characterized by two steps, i.e., the "E" step and the "M" step. HMM is a tool designed for representing probability distributions of sequences of observations. It can be considered a generalization of a mixture-based model, where the hidden variables, which control the specific mixture of the component to be selected for each observation, are related to

each other through a Markov process, rather than being independent of each other.

Applications

In the following, some applications of supervised learning aided network association algorithms are elaborated, which are also summarized in Table 1. KNN algorithms are beneficial for traffic prediction, for anomaly detection, as well as for modulation classification. Specifically, in order to capture the dynamic characteristics of radio resource demands, a weighted KNN model was proposed based on a large-scale historical data set from cellular operator networks, which was used to explore both temporal and spatial characteristics of radio resource margins to enhance the load balance (Feng et al. 2017). Moreover, the authors of Onireti et al. (2016) proposed a KNN-based anomaly detecting algorithms for improving cell outage detection accuracy. In Aslam et al. (2012), genetic programming and KNN were combined in order to improve the modulation classification accuracy, which provided a reliable modulation classification scheme for the secondary users (SU) in cognitive radio networks.

Furthermore, the SVM-aided learning models can be used for estimating radio parameters, for predicting mobile terminal’s usage pattern, and for guiding channel selection. For example, in order to generalize the SVM function for employment in data classification problems, its hierarchical version referred to as H-SVM was proposed in Feng and Chang (2012), where each hierarchical level consisted of

a finite number of SVM classifiers. This regime was used for the estimation of the Gaussian channel’s noise level in a MIMO-aided wireless network. By exploiting the training data, the H-SVM model was trained for the estimation of the channel noise statistics. SVM can also be used for learning the mobile terminal’s specific usage pattern in diverse spatiotemporal and device contexts, as discussed in Donohoo et al. (2014). This may then be exploited for prediction of the configuration to be used in the location-specific interface for HetNets constituted by diverse cells. In Thilina et al. (2016), the common control channel for SUs during a given frame was selected by employing an SVM-based learning technique for a cognitive radio network, which was capable of implicitly learning the surrounding environment cooperatively in an online fashion.

The Bayesian learning model may be readily invoked for spectral characteristic learning and estimation. The authors of Wen et al. (2015) estimated both the channel parameters of the desired links in a target cell and those of the interfering links of the adjacent cells, relying on the sparse Bayesian learning techniques, where the channel component was first modeled by a GM, and then estimated with the aid of the EM algorithm. In Choi and Hossain (2013), a cooperative wideband spectrum sensing scheme based on the EM algorithm was proposed for the detection of a primary user (PU) supported by a multi-antenna assisted cognitive radio net-

Machine Learning Paradigms in Wireless Network Association, Table 1 Applications of supervised learning aided network association algorithms

Method	Scenario	Applications	Reference
KNN	Cognitive radio	Radio resource capture	(Feng et al. 2017)
		Modulation classification	(Aslam et al. 2012)
	HetNets	Anomaly detection	(Onireti et al. 2016)
SVM	MIMO	Channel noise estimation	(Feng and Chang 2012)
	Cognitive radio	Channel selection	(Thilina et al. 2016)
	HetNets	Usage pattern prediction	(Donohoo et al. 2014)
Bayesian learning	MIMO	Channel parameter estimation	(Wen et al. 2015)
	Cognitive radio	Spectrum sensing	(Choi and Hossain 2013)
		Signal estimation and detection	(Assra et al. 2016)

work. Furthermore, in Assra et al. (2016), the authors constructed a HMM relying on a two-state hidden Markov process, where the PUs are present or absent, and a two-state observation space, indicating whether the PUs are present or absent. Furthermore, the EM algorithm was invoked for finding the true channel parameters, such as the sojourn time of the available channels, the inactive states of the PUs, as well as the PUs' signal strength.

Unsupervised Learning in Network Association

Models

K-means clustering aims for partitioning n observations into k clusters, where each observation belongs to the closest cluster. It defines the centroid of a cluster as the center of gravity, i.e., the mean value of the points within the cluster. The clustering algorithm proceeds in an iterative manner, where an object is assigned to the specific cluster whose centroid is nearest to the object based on the Euclidean distance, and then the in-cluster differences are minimized by iteratively updating the cluster centroid, until convergence is achieved.

Principal component analysis (PCA) transforms a set of potentially correlated variables into a set of uncorrelated variables referred to as the principal components, where the number of principal components is less than or equal to the number of original variables. Basically, the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The principal components are orthogonal, because they are the eigenvectors of the covariance matrix, which is symmetric. By contrast, *independent component analysis* (ICA) is a statistical technique conceived for revealing hidden factors that underlie sets of random variables, measurements, or signals. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables

are assumed to be non-Gaussian and mutually independent, and they are referred to as the independent components of the observed data, which can be found by ICA.

Applications

Table 2 summarizes some applications relying on unsupervised learning assisted network association schemes. To elaborate a little further, clustering is a common problem in wireless networks, especially in heterogeneous scenarios associated with diverse cell sizes as well as Wi-Fi and D2D networks. For example, the small cells have to be carefully clustered for avoiding interference using coordinated multipoint transmission, while the mobile users are clustered for obeying an optimal offloading policy; the devices are clustered in D2D networks for the sake of achieving high energy efficiency; the Wi-Fi users are clustered for maintaining an optimal access point association, etc. A mixed integer programming problem was formulated for jointly optimizing both the gateway partitioning and the virtual channel allocation based on classic k -means clustering, which was employed for partitioning the mesh access points into several groups in a hybrid optical/wireless network scenario for the sake of reducing the overall wireless tele-traffic by encouraging the utilization of the high-capacity optical infrastructure (Xia et al. 2012). In Hajjar et al. (2017), a K -means-based relay selection algorithm was proposed for creating low-power small cells in an LTE macro cell within a multi-cell scenario, where the network's total capacity was increased without the need of additional infrastructure.

Both the PCA and ICA constitute powerful statistical signal processing techniques devised for recovering statistically independent source signals from their linear mixtures. One of their major applications may be found in the area of anomaly-, fault-, and intrusion-detection problems of wireless networks, which rely on traffic monitoring. Furthermore, similar problems may also be solved in sensor networks, mesh networks, etc. They can also be invoked for the physical layer signal dimension reduction of massive MIMO systems or for classifying

Machine Learning Paradigms in Wireless Network Association, Table 2 Applications of unsupervised learning aided network association algorithms

Method	Scenario	Applications	Reference
K-means clustering	Hybrid network	Virtual channel allocation	(Xia et al. 2012)
	LTE network	Relay selection	(Hajjar et al. 2017)
	MIMO	Spectral efficiency	(Liang et al. 2016)
	Optical networks	Signal detection	(Zhao et al. 2006)
	WSN	Optimal sensor deployment	(Ateş et al. 2017)
PCA	Wi-Fi	User detection	(Zhu et al. 2017)
	WSN	Efficient localization	(Li et al. 2017)
ICA	Cognitive radio	Signal sources detection	(Nguyen et al. 2013)
	Smart grid	Energy efficiency	(Qiu et al. 2011)

the primary users' behaviors in cognitive radio networks. In Qiu et al. (2011), PCA and ICA were applied in a smart grid scenario for recovering the simultaneous wireless transmissions of smart utility meters installed in each home, which were capable of enhancing both the transmission efficiency by avoiding the channel estimation in each frame and the data security by eliminating any wideband interference or jamming signals. Another pertinent example is found in cognitive radio scenarios, where the so-called Boolean ICA relied on the Boolean mixing of OR, XOR, and other functions of binary signals (Nguyen et al. 2013), where an iterative algorithm, termed as the binary ICA, was developed for determining the activities of the underlying latent signal sources, such as the PUs. It was demonstrated that given m monitors or SUs, the activities of up to $(2m - 1)$ distinct PUs can be inferred.

Reinforcement Learning in Network Association

Models

Multiarmed bandit relies on a player at a row of slot machines who has to decide which machines to play on and how many times to play on each. He/she will receive a random reward provided by the selected machine. The objective of the game is to maximize the sum of rewards earned through a sequence of lever pulls. It has become one of the popular examples of sequential decision-making

problems striking an exploration-exploitation trade-off relying on limited information.

Markov decision process (MDP) provides a mathematical framework for modeling decision-making in specific situations, where the outcomes are partly random and partly under the control of a decision-maker. At each time step, the process is in some state S , and the decision-maker may opt for any of the legitimate actions A that is available in state S . The process responds at the next time step by randomly moving into a new state S' and giving the decision-maker a corresponding reward $r(S, A)$. The probability that the process moves into its new state S' is influenced both by the specific action chosen and by the system's inherent transitions. As an important member of MDP family, *partially observable Markov decision process* (POMDP) may be suitable for a general scenario, where the agent is unable to directly observe the underlying state transitions and hence only has partial knowledge. The agent has to keep track of both the probability distribution of the legitimate states, based on a set of observations, and the observation probabilities and of the underlying MDP.

Q-learning may be invoked for finding an optimal action policy for any given finite Markov decision process, especially when the system model is unknown. It is a model-free reinforcement learning technique, and as such it can be used in conjunction with MDP models. In such a case, the Q-learning model is also comprised of an agent, of the states, and of a set of actions per state. By executing an action in a

Machine Learning Paradigms in Wireless Network Association, Table 3 Applications of reinforcement learning aided network association algorithms

Method	Scenario	Applications	Reference
Multiarmed bandit	D2D network	Channel selection	(Maghsudi and Stańczak 2015)
	Li-Fi network	AP selection	(Wang et al. 2018)
MDP/POMDP	WSN	Power control	(Aprem et al. 2013)
	Super Wi-Fi	Energy harvesting	(Wang et al. 2016)
	D2D	Network coding	(Karim et al. 2017)
Q-learning	Femtocell network	Resource allocation	(Alnwaimi et al. 2015)
	HetNets	Admission control	(Chen et al. 2009)
	Dense cellular network	Traffic offloading	(Fakhfakh and Hamouda 2017)

specific state, the agent gleans a reward and the goal is to maximize its accumulated reward. Such a reward is illustrated by a Q -function, where “ Q ” is initialized to be a fixed value. Then, “ Q ” is updated in an iterative manner after the agent carries out an action and observes the resultant reward as well as the associated new state at each time instant.

Applications

Table 3 lists a range of applications of reinforcement learning assisted network association schemes. Generally, multiarmed bandit may be beneficially used in multiplayer adaptive decision-making problems, where selfish players infer an optimal joint action profile from their successive interactions with a dynamic environment and finally settle at some equilibrium point. This problem has indeed been encountered in many wireless networking scenarios, with a compelling one being the channel selection problem (Maghsudi and Stańczak 2015) and another one in the context of access point (AP) selection problem (Wang et al. 2018). Specifically, every mobile user was modeled as a player of the multiarmed bandit game, while the channels/APs were regarded as arms and choosing a channel or an AP corresponds to pulling an arm. The authors of Maghsudi and Stańczak (2015) proposed a channel selection strategy consisting of two main blocks for D2D communication system integrated into a cellular network, namely, the calibrated forecasting and the no-regret bandit learning strategies. Moreover, the authors of Maghsudi

and Stańczak (2015) proposed a multiarmed bandit-aided AP selection algorithm for a visible light communication (VLC) and Wi-Fi hybrid system (Li-Fi).

The family of MDP/POMDP models constitutes ideal tools for supporting decision-making in wireless networks, where the users may be regarded as agents and the network constitutes the environment. For instance, in Aprem et al. (2013), the transmission power control problems in energy harvesting wireless sensor networks (WSN) were investigated using the POMDP model, where the state space was defined by including the battery state, the channel state, the packet transmission/reception states, and an action by the node, which corresponded to sending a packet at a certain power level. In Wang et al. (2016), a POMDP-aided AP section algorithm was proposed for improving the energy efficiency of the super Wi-Fi system, including a low-complexity energy function-based solution.

Q-learning has also been extensively applied in HetNets, usually in conjunction with the aforementioned MDP models. In Alnwaimi et al. (2015), the authors presented a heterogeneous fully distributed multi-objective strategy based on a reinforcement learning model constructed for the self-configuration/optimization of femtocells. The model was supposed to solve both the resource allocation and interference coordination problem in the downlink of femtocell networks under carefully constructed restrictions for avoiding interference and for meeting the quality of service (QoS) requirements. Furthermore,

in Chen et al. (2009), a fuzzy Q-learning admission control for wideband code division multiple access (WCDMA)/wireless local area network (WLAN) HetNets was proposed, which considered not only QoS requirements but also multiple system measures, such as interferences, the numbers of real-time and non-real-time users, user’s mobility, etc.

Deep Learning in Network Association

Models

Artificial neural networks are a set of algorithms inspired by the biological neural networks that constitute animal brains, which help us cluster and classify. The *deep neural network* (DNN) is a kind of artificial neural network, with multiple layers between the input and output layers, which can model complex nonlinear relationships. The layers are made of nodes, where a node is just a place where computation happens, loosely patterned on a neuron in the human brain. A node combines input from the data with a set of weights that either amplify or dampen that input, thereby assigning significance to inputs for the task considered. The nodes residing in “deeper” layer of the neural net are capable of recognizing more complex features, since they aggregate and recombine features from the previous layer. In other words, the extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network.

The *recurrent neural network* (RNN) is another class of artificial neural network where

connections between units form a directed cycle, which allows it to exhibit dynamic temporal behavior. Unlike traditional neural networks where all inputs are independent of each other, RNN is called “recurrent” because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Hence, RNN can be viewed to have a “memory” which captures information about what has been calculated so far. RNNs can use their internal memory to process arbitrary sequences of inputs.

Applications

In Table 4, we summarize some typical applications of deep learning aided network association algorithms. DNN algorithms are readily used for traffic control, wireless localization, activity recognition, resource allocation, and anomaly detection. Specifically, in order to improve traffic control in HetNets, the authors in Kato et al. (2017) first characterized the HetNet traffic and then proposed a supervised DNN system, which exploited the extracted characterizations to achieve excellent signaling overhead, throughput, and delay. Moreover, in a cognitive radio network, the DNN-based algorithm was proposed to learn features that indicated the influence of the target on the wireless signals, which substantially improved wireless localization and activity recognition (Wang et al. 2017a).

Furthermore, RNN can be applied to traffic classification, interference cancellation, and channel equalization. For instance, in Lopez-Martin et al. (2017), a combination model of a RNN and a convolutional neural network (CNN) was proposed to classify both the traffic and the

Machine Learning Paradigms in Wireless Network Association, Table 4

Applications of deep learning aided network association algorithms

Method	Scenario	Applications	Reference
DNN	HetNets	Traffic control	(Kato et al. 2017)
	Cognitive radio	Wireless localization	(Wang et al. 2017a)
	Smart grid	Attack detection	(He et al. 2017)
	WSN	Sensor calibration	(Wang et al. 2017b)
RNN	IoT network	Traffic classification	(Lopez-Martin et al. 2017)
	MIMO	Interference cancellation	(Mostafa 2017)
		Channel equalization	(Zhao et al. 2011)
	Cellular network	Power control	(Chen et al. 2006)

behavior of heterogeneous devices and services in IoT networks, which provided a superior detection result without requiring any feature engineering. Besides, as discussed in Mostafa (2017), RNN was beneficially invoked for interference cancellation in MIMO systems. The proposed RNN algorithm showed good performance in terms of suppressing impulsive noise while assuring the Lyapunov stability simultaneously.

Conclusions

This entry reviewed the benefits of artificial intelligence-aided wireless systems equipped with machine learning. We introduced the major families of machine learning algorithms and discussed their applications in the context of massive MIMOs, the smart grid, cognitive radios, HetNets, small cells, D2D networks, etc. The classes of supervised, unsupervised, reinforcement, and deep learning tools were investigated, along with the corresponding modeling methodology and possible future applications in wireless network association. In a nutshell, machine learning is an exciting area for artificial intelligence-aided networking research!

References

- Alnwaimi G, Vahid S, Moessner K (2015) Dynamic heterogeneous learning games for opportunistic access in LTE-based macro/femtocell deployments. *IEEE Trans Wirel Commun* 14(4):2294–2308
- Alsheikh MA, Lin S, Niyato D, Tan HP (2014) Machine learning in wireless sensor networks: algorithms, strategies, and applications. *IEEE Commun Surv Tutor* 16(4):1996–2018
- Aprem A, Murthy CR, Mehta NB (2013) Transmit power control policies for energy harvesting sensors with retransmissions. *IEEE J Sel Top Sign Proces* 7(5):895–906
- Aslam MW, Zhu Z, Nandi AK (2012) Automatic modulation classification using combination of genetic programming and KNN. *IEEE Trans Wirel Commun* 11(8):2742–2750
- Assra A, Yang J, Champagne B (2016) An EM approach for cooperative spectrum sensing in multiantenna CR networks. *IEEE Trans Veh Technol* 65(3):1229–1243
- Ateş E, Kalayci TE, Uğur A (2017) Area-priority-based sensor deployment optimisation with priority estimation using K-means. *IET Commun* 11(7):1082–1090
- Chen YS, Chang CJ, Hsieh YL (2006) A channel effect prediction-based power control scheme using PRNN/ERLS for uplinks in DS-CDMA cellular mobile systems. *IEEE Trans Wirel Commun* 5(1):23–27
- Chen YH, Chang CJ, Huang CY (2009) Fuzzy Q-learning admission control for WCDMA/WLAN heterogeneous networks with multimedia traffic. *IEEE Trans Mob Comput* 8(11):1469–1479
- Choi KW, Hossain E (2013) Estimation of primary user parameters in cognitive radio systems via hidden Markov model. *IEEE Trans Signal Process* 61(3):782–795
- Clancy C, Hecker J, Stuntebeck E, O'Shea T (2007) Applications of machine learning to cognitive radio networks. *IEEE Wirel Commun* 14(4):47–52
- Donohoo BK, Ohlsen C, Pasricha S, Xiang Y, Anderson C (2014) Context-aware energy enhancements for smart mobile devices. *IEEE Trans Mob Comput* 13(8):1720–1732
- Fakhfakh E, Hamouda S (2017) Optimised Q-learning for WiFi offloading in dense cellular networks. *IET Commun* 11(15):2380–2385
- Feng VS, Chang SY (2012) Determination of wireless networks parameters through parallel hierarchical support vector machines. *IEEE Trans Parallel Distrib Syst* 23(3):505–512
- Feng Z, Li X, Zhang Q, Li W (2017) Proactive radio resource optimization with margin prediction: a data mining approach. *IEEE Trans Veh Technol* 66(10):9050–9060
- Hajjar M, Aldabbagh G, Dimitriou N, Win MZ (2017) Hybrid clustering scheme for relaying in multi-cell LTE high user density networks. *IEEE Access* 5:4431–4438
- He Y, Mendis GJ, Wei J (2017) Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism. *IEEE Trans Smart Grid* 8(5):2505–2516
- Jiang C, Zhang H, Ren Y, Han Z, Chen KC, Hanzo L (2017) Machine learning paradigms for next-generation wireless networks. *IEEE Wirel Commun* 24(2):98–105
- Karim MS, Sorour S, Sadeghi P (2017) Network coding for video distortion reduction in device-to-device communications. *IEEE Trans Veh Technol* 66(6):4898–4913
- Kato N, Fadlullah ZM, Mao B, Tang F, Akashi O, Inoue T, Mizutani K (2017) The deep learning vision for heterogeneous network traffic control: proposal, challenges, and future perspective. *IEEE Wirel Commun* 24(3):146–153
- Li X, Ding S, Li Y (2017) Outlier suppression via non-convex robust PCA for efficient localization in wireless sensor networks. *IEEE Sensors J* 17(21):7053–7063
- Liang HW, Chung WH, Kuo SY (2016) Coding-aided K-means clustering blind transceiver for space shift keying MIMO systems. *IEEE Trans Wirel Commun* 15(1):103–115

- Lopez-Martin M, Carro B, Sanchez-Esguevillas A, Lloret J (2017) Network traffic classifier with convolutional and recurrent neural networks for internet of things. *IEEE Access* 5:18042–18050
- Maghsudi S, Stańczak S (2015) Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. *IEEE Trans Wirel Commun* 14(3):1309–1322
- Mostafa M (2017) Stability proof of iterative interference cancellation for OFDM signals with blanking nonlinearity in impulsive noise channels. *IEEE Signal Process Lett* 24(2):201–205
- Nguyen H, Zheng G, Zheng R, Han Z (2013) Binary inference for primary user separation in cognitive radio networks. *IEEE Trans Wirel Commun* 12(4):1532–1542
- Onireti O, Zoha A, Moysen J, Imran A, Giupponi L, Imran MA, Abu-Dayya A (2016) A cell outage management framework for dense heterogeneous networks. *IEEE Trans Veh Technol* 65(4):2097–2113
- Qiu RC, Hu Z, Chen Z, Guo N, Ranganathan R, Hou S, Zheng G (2011) Cognitive radio network for the smart grid: experimental system architecture, control algorithms, security, and microgrid testbed. *IEEE Trans Smart Grid* 2(4):724–740
- Thilina KGM, Hossain E, Kim DI (2016) DCCC-MAC: a dynamic common-control-channel-based MAC protocol for cellular cognitive radio networks. *IEEE Trans Veh Technol* 65(5):3597–3613
- Wang J, Jiang C, Han Z, Ren Y, Hanzo L (2016) Network association strategies for an energy harvesting aided super-wifi network relying on measured solar activity. *IEEE J Sel Areas Commun* 34(12):3785–3797
- Wang J, Zhang X, Gao Q, Yue H, Wang H (2017a) Device-free wireless localization and activity recognition: a deep learning approach. *IEEE Trans Veh Technol* 66(7):6258–6267
- Wang Y, Yang A, Chen X, Wang P, Wang Y, Yang H (2017b) A deep learning approach for blind drift calibration of sensor networks. *IEEE Sensors J* 17(13):4158–4171
- Wang J, Jiang C, Zhang H, Zhang X, Leung VC, Hanzo L (2018) Learning-aided network association for hybrid indoor LiFi-WiFi systems. *IEEE Trans Veh Technol* 67(4):3561–3574
- Wen CK, Jin S, Wong KK, Chen JC, Ting P (2015) Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning. *IEEE Trans Wirel Commun* 14(3):1356–1368
- Xia M, Owada Y, Inoue M, Harai H (2012) Optical and wireless hybrid access networks: design and optimization. *J Opt Commun Netw* 4(10):749–759
- Zhao T, Nehorai A, Porat B (2006) K-means clustering-based data detection and symbol-timing recovery for burst-mode optical receiver. *IEEE Trans Commun* 54(8):1492–1501
- Zhao H, Zeng X, Zhang J, Li T (2011) Equalisation of non-linear time-varying channels using a pipelined decision feedback recurrent neural network filter in wireless communication systems. *IET Commun* 5(3):381–395
- Zhu H, Xiao F, Sun L, Wang R, Yang P (2017) R-TTWD: robust device-free through-the-wall detection of moving human with WiFi. *IEEE J Sel Areas Commun* 35(5):1090–1103

Machine to Machine (M2M)

► Machine-Type Communication

Machine-Learning (ML) in 4G Networks

► Survey on Machine-Learning Techniques in LTE Networks

Machine-Type Communication

Chia-Peng Lee¹ and Phone Lin²

¹Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

²Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

Synonyms

[Internet of things](#); [Machine to machine \(M2M\)](#)

Definition

The machine-type communication (MTC) is a new type of communication technologies for the machine-to-machine (M2M) or Internet of Things (IoTs) with few or without human intervention (3GPP 2014), and it is proposed by the Third Generation Partnership Project (3GPP) working group.

Historical Background

The standardization work for the MTC was started by 3GPP TSG SA1 working group in early 2005. In 2007, a technical report on facilitating M2M communication in the 3GPP systems was exported (3GPP 2007). The 3GPP also identified the service requirements and system improvements for the MTC in release 10 (3GPP 2014), and a specific protocol implementation is described in release 11 (3GPP 2012).

The features of MTC specified by the 3GPP are listed as follows (3GPP 2014):

- Low mobility: The MTC devices have low mobility or move only within a specific area.
- Time controlled: The MTC devices transmit the data during specific time intervals.
- Small data transmissions: The applications have a small amount of data transmission or reception.
- Mobile originated only (or infrequent mobile terminated): Most of application calls are initiated by the MTC device.
- MTC monitoring: Most of applications monitor the behaviors and the events of the MTC devices.
- Priority alarm: The alerting function in applications reports anomaly events, e.g., the MTC devices were stolen or destroyed.
- Group-based MTC features: The system is optimized with grouped MTC devices.

Figure 1 shows the architecture of the MTC in release 11 (3GPP 2012, 2018), where an MTC device connects to the LTE core network through the LTE-Uu air interface. The network entities and interfaces are described as follows:

- The MTC Interworking Function (MTC-IWF) is a network element that receives external application service signaling through the Tsp interface.
- The Service Capability Server (SCS) communicates with MTC devices through the MTC-IWF. The MTC devices can access one or more application services controlled and

managed by the Application Servers (ASs) through the SCS.

- The S6m interface exercises between the MTC-IWF and the Home Subscriber System (HSS). The interface is used by the MTC-IWF to obtain the IDs of an MTC device from the HSS, which include the International Mobile Subscriber Identity (IMSI) and Mobile Subscriber Integrated Services Digital Network Number (MSISDN).
- The MTC-IWF sends trigger messages to the Short Message Service Center (SMS-SC) through the T4 interface.

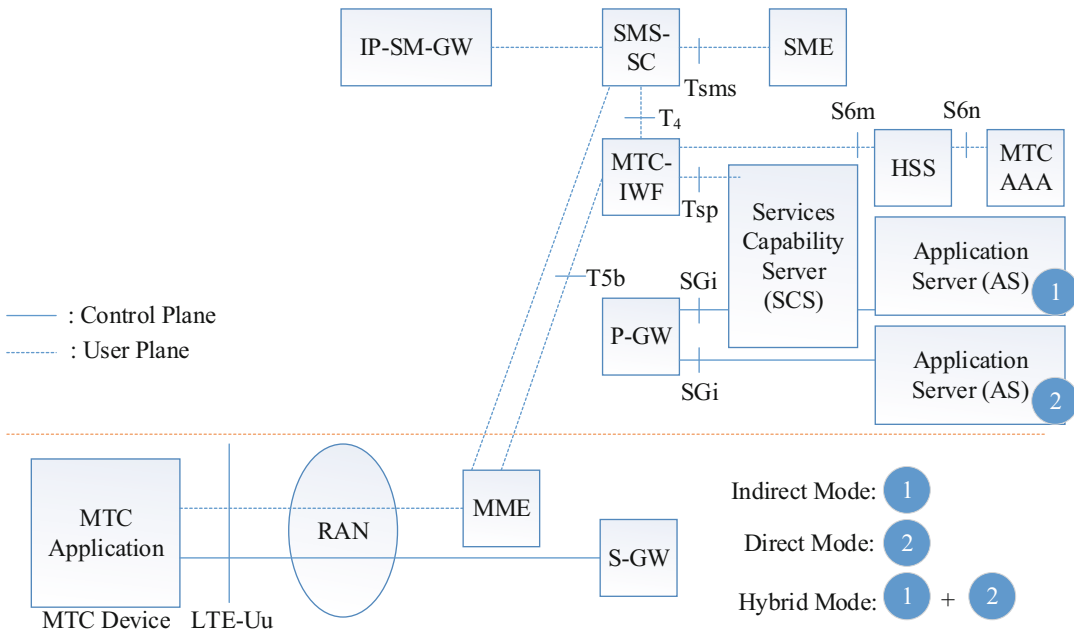
Based on the above network architecture, the 3GPP defines three modes for the core network to connect to the MTC ASs, including the indirect mode, the direct mode, and the hybrid mode. The executions of the three connected modes are described below:

- Indirect mode: The AS connects to the core network through the SCS.
- Direct mode: The AS connects the core network through the PDN Gateway (P-GW) that is a router in the core network and provides IP connectivity between the core network and external data network.
- Hybrid mode: The LTE network operator can provide both of indirect and direct modes.

Key Applications

The 3GPP release 14 (3GPP 2017) lists the key applications that fall into the following service areas.

- The following MTC applications belong to the security service area:
 1. An alarm system includes a set of MTC devices that can trigger an alarm. Examples of the MTC devices can be gas sensors, temperature sensors, etc.
 2. A surveillance system uses the MTC devices (e.g., camera, closed-circuit



Machine-Type Communication, Fig. 1 The architecture of LTE MTC network

television; CCTV) to observe or monitor users' behaviors and activities.

3. The backup-for-landline system is usually used for home security, which is equipped with the MTC devices to monitor the land-line. If the landline is broken, the MTC devices will send a warning message to the monitor system and notify that the landline has something wrong and needs to fix.
 4. The physical access control system accommodates a set of physical equipment (e.g., RFID door and gate controller). The physical access control system is implemented in a building and authorizes legal access.
 5. With the car/driver security system, in a connected vehicle, the MTC devices can retrieve the information (e.g., revolution per minute; RPM), based on which, the system can diagnose whether the car is safe or not. It also ensures the safety of the driver.
- The following applications fall into the tracking and tracing service area:
 1. With the fleet management system, a car installed with the MTC devices which combine GPS and short message service to keep track immediate location and history trajectory of a car.
 2. In the pay-as-you-drive service, the car insurance fees are easy to collect based on the information in car such as speed, history trajectory, and driving time, and then the collected data transmit to the insurance company.
 3. With the asset tracking system, every asset installs an MTC device with GPS service, and the MTC device sends the location to the asset tracking system for easy tracking the location of the asset.
 4. In the navigation service, the drivers can change the driving path in advance through the geographical information and other information reported by the MTC devices to avoid the traffic jams.
 5. With traffic information system, the car equips with MTC devices combined GPS function that reports location information

- to the MTC server. Then, based on the reported location, the MTC server provides the traffic information of that of reported location to inform driver to avoid the traffic jam.
6. With the road tolling system, the roadside units (RSUs) are deployed along the free-way. When a car equipped with the MTC device passes by that of the freeway, the RSU senses the equipped MTC device and then makes a charge.
 7. With the traffic optimization/steering system, based on the location collected by MTC device in a car, it is easy to provide the guideline to the driver to achieve the traffic optimization and steering.
- The following MTC applications belong to the payment service area:
 1. With the point-of-sales system, the MTC device reads commodity information (e.g., good name, price, and qualify) and then sends the information to relevant departments through communication networks for analysis to improve operating efficiency of the system.
 2. The vending machine system is equipped with MTC devices to inform the sales that the items have been sold out by email or short messages, and they should replenish the items immediately. Another application is if the vending machine is broken down, the MTC device is triggered to send an alert (e.g., short message) to the sales.
 - The following applications fall into the health service area:
 1. With the monitoring vital signs system, the person wears MTC devices that monitor the changing of vital signs (e.g., the pulse, temperature, breathing, and blood pressure). If the monitoring vital signs change, the MTC devices send an alert message to the monitoring vital signs system.
 2. In the aged and handicapped system, the MTC devices are deployed in the home environment to keep track of the behavior of elderly people.
 3. With the remote health diagnostics system, the doctor can remotely monitor the health status and chronic conditions of patients anywhere anytime through the MTC devices and make reaction immediately when an emergency event occurs.
 - The following applications fall into the remote maintenance and control service area:
 1. In the programmable logic controllers (PLCs) system, MTC devices can control PLC by predefined instructions (e.g., turning on the light at home) which are stored in PLCs memory.
 2. In the sensor system, the MTC devices are deployed as sensors (e.g., power, fire, smoke, gas, water, wind, etc.) to monitor the environment.
 3. In the lighting system, the MTC devices can detect environmental change (e.g., from day to night) and send a notification to the lighting system to turn on the light.
 4. In the vending machine control system, the vending machine is installed with MTC devices to determine whether the vending machine operates normally or abnormally. If the detected result is abnormal, the MTC devices send a warning message to the vending machine control system.
 5. In the vehicle diagnostics system, the MTC devices are installed in a connected car to collect the information from the car, and then send the information to the vehicle diagnostics system, based on which the diagnostics system analyzes whether the car is in the safe status or the dangerous status.
 - The following MTC applications belong to the metering service area:
 1. In the gas system or the water system, the MTC devices are installed as smart meters,

which collect the information about the usage of gas or water. The information is referenced to charge the users for the usage of gas or water.

2. In the heating system, the MTC devices are deployed as sensor to sense the ambient temperature to control the heating automatically.
3. In the grid control system, the MTC devices collect information of power supply status in the home environment, and send the information to the grid control system to adjust the production and transmission of electricity to reduce the power consumption for power saving.

Cross-References

- [Internet of Things](#)
- [Machine to Machine \(M2M\)](#)

References

- 3 GPP (2007) Study on facilitating machine to machine communication in 3 GPP systems: TR 22.868
- 3GPP (2012) System improvements for Machine-Type Communications (MTC):TR 23.888
- 3GPP (2014) Service requirements for Machine-Type Communications (MTC): TS 22.368
- 3GPP (2017) Service requirements for Machine-Type Communications (MTC): TS 22.368
- 3GPP (2018) Architecture enhancements to facilitate communications with packet data networks and applications: TS 23.682

Recommended Reading

- 3GPP (2013) Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements: TR 23.887
- Jain P et al (2012) Machine type communications in 3GPP systems. *IEEE Commun Mag* 50(11):28–35. <https://doi.org/10.1109/mcom.2012.6353679>
- Lien S-Y et al (2011) Toward ubiquitous massive accesses in 3 GPP machine-to-machine communications. *IEEE Commun Mag* 49(4):66–74. <https://doi.org/10.1109/mcom.2011.5741148>
- Taleb T, Kunz A (2012) Machine type communications in 3GPP networks: potential, challenges, and solutions. *IEEE Commun Mag* 50(3):178–184. <https://doi.org/10.1109/mcom.2012.6163599>

Macrocell–Small Cell Systems

- [Resource Management in Macrocell–Small Cell Systems and D2D-Assisted Cellular Systems](#)

Malware

- [New Variants of Mirai and Analysis](#)

Maritime Internet of Vessels

Wang Zhen¹ and Bin Lin²

¹Dalian Neusoft University of Information,
Dalian Maritime University, Dalian, China

²School of Information Technology, Dalian
Maritime University, Dalian, China

Synonyms

[Internet of Things for vessels](#); [Networking of marine vessels](#)

Definitions

Maritime Internet of Vessels is a kind of information service network for intelligent shipping which integrates Internet of Things technology. It takes vessels, waterways, and land-shore facilities as basic nodes and information sources and connects vessels through wireless communication technology, aiming at realizing the refinement of shipping management, the comprehensiveness of industry services, and the humanization of user experience.

Key Points

Historical Background

With the development of shipping industry, the amount of data and information acquired from

ships, cargo, operators, and navigation environment is increasing at the speed of TB and PB, and the traditional maritime communication systems can hardly afford for the development requirements at present. How to transmit, store, and manage the massive data effectively has become a research hotspot. At present, more than 80% of the world's trade is completed by shipping, and more than 50,000 merchant ships are engaged in international trade, maintaining the stability of global material transport. However, accidents related to navigation continue to occur despite the development and availability of a number of ship- and shore-based technologies that promise to improve situational awareness and decision-making. In 2018, there were 255 pirate attacks and harassment incidents worldwide, and 126 crew members were kidnapped by pirates worldwide. Safety, efficiency, and environmental protection of maritime communications have always been the core issues in the development of intelligent shipping, as well as the key points in ship collision avoidance, navigation assistance, and maritime positioning.

In recent years, the “intelligent” manufacturing model and products have extended to various fields, and many high tech and equipment have been greatly developed and widely used in the field of navigation, especially on board ships. “Intelligent shipping” and “e-Navigation” are respectively promoted on two independent main lines of development to jointly realize the safety, efficiency, and environmental protection of maritime shipping. In addition, the Maritime Internet of Vessels will bring shipping into a new era in which vessels and vessels, vessels and waterways, and vessels and people are interconnected and the original risk-ridden and uncertain voyages begin to become intelligent and controllable. To better comprehend the development of the Maritime Internet of Vessels, this entry makes a detailed investigation about the existing networking modes for Maritime Internet of Vessels and puts forward the further development. With different kinds of possible network access and interconnection, the ubiquitous links between vessels and vessels and people can be realized; the intelligent perception, recognition,

and management of vessels and processes can also come true; and the maritime navigation will become more intelligent, controllable, safe, and environmentally friendly.

The Existing Networking Modes for Maritime Internet of Vessels

At the 81st meeting of the Maritime Safety Committee (MSC) of the International Maritime Organization (IMO), an e-Navigation proposal was put forward jointly in order to integrate and harmonize the maritime communication systems and better exchange and unify information between shore-based and shipboard users through electronic means. The IMO has defined e-Navigation as “the harmonized collection, integration, exchange, presentation and analysis of marine information on board and ashore by electronic means to enhance berth to berth navigation and related services for safety and security at sea and protection of the marine environment” (Weinrit and Weinrit 2011). The establishment of e-Navigation makes the information exchange between ship and ship and ship and shore safer, more convenient, and more efficient. On this basis, the concept of Maritime Service Portfolios (MSP) came into being which mainly refers to a set of standardized, operational, and technical maritime service information provided by the coastal side to navigators in a given sea area, waterway, port, and other similar areas. According to different applications and services, the networking modes for Maritime Internet of Vessels will be described from the perspective of conventional communications and emergency communications:

- **The conventional communications**

The conventional communications are mainly used for maritime applications such as electronic chart display and information system (ECDIS), normal navigation, voyage planning, crew and passengers' communications, etc. These systems mainly operate at medium- and high-frequency systems and very-high-frequency (MF, HF, VHF) band to achieve a certain degree of voice and data communication.

- (1) Digital selective calling (DSC) (MF/HF & VHF bands)

Digital selective calling or DSC is a standard for sending pre-defined digital messages to another station or group of stations for distress or general communications via the medium-frequency (MF), high-frequency (HF), and very-high-frequency (VHF) maritime radio systems (de Sousa and Pelas 2011). The International Radio Advisory Committee (CCIR) has allocated a DSC channel in the 2 MHz band of MF; the 4, 6, 8, 12, and 16 MHz band of HF; and the VHF marine channel 70 (156.525 MHz) for the DSC distress call (alarm) frequency and the radio telephone, radio distress, and secure communication frequency.

- (2) Narrowband direct printing (NBDP or radio telex, MF)

NBDP is a technology that automates radio signals to telegraphy which has been used for ship location reporting and the issuance of weather warnings and forecasts at coastal stations, while the use of NBDP for general communications is decreasing. As an integral part of GMDSS, NBDP is a FSK modulated to 0.5 kHz HF channel, which supports low-speed data transmission (100 bps) over 1.6–26.5 MHz for maritime mobile services. It can be used as text-based distress tracking communication and general communication between ship-to-ship, ship-to-shore, and shore-to-ship, especially to overcome language difficulties.

- (3) Navigational Telex (NAVTEX, MF)

NAVTEX (Navigational Telex), as a major element of the global maritime distress and safety system, is diffusely used for delivery of navigational and meteorological warnings and forecasts, as well as urgent maritime safety information to ships. It mainly provides a low-cost, simple, and automated means of receiving this information aboard ships

at sea within approximately 370 km off shore. It adopts FSK modulation scheme and broadcasts messages in English on MF band of 518 kHz, while 490 and 4209.5 kHz are used to broadcast in English and/or local language.

- (4) Automatic identification system (AIS, VHF)

The automatic identification system (AIS) is an automatic tracking system aimed at improving the safety of navigation by assisting ships in effective navigation, environmental protection, and operation of vessel traffic services (VTS). It provides a means for ships to electronically exchange ship data including identification, position, course, speed and navigational status with other nearby ships and AIS base stations, and satellites. AIS uses VHF channels and adopts a technology called self-organized time division multiple access (SOTDMA) to ensure that the VHF transmissions of different transceivers do not occur at the same time. Besides, AIS is also integrated into other maritime devices such as aids to navigation (A to N), search and rescue transmitter, and man overboard units (Lázaro et al. 2017).

- (5) VHF data exchange system (VDES, VHF)

VDES is an enhanced and upgraded system of AIS for alleviating the pressure in the existing frequency band of the AIS system and ensuring the implementation of the normal function in the field of maritime mobile services. On the basis of integrating existing AIS functions, VDES adds special application message (ASM) and broadband VHF data exchange (VDE) functions, which can effectively alleviate the pressure of existing AIS data communication and provide effective auxiliary means for the protection of ship navigation safety. At the same time, it will comprehensively improve the ability and frequency efficiency of marine data communication, which is of great significance to promote the development

of marine radio digital communication industry. Furthermore, the standardization process of VDES is quite mature, and it is expected to become an ITU standard soon (Lázaro et al. 2017).

- The emergency communications

Besides general communications, the emergency communications also focus on security issues of maritime applications which are related to the safety of life and property at sea, distress and urgency alerting/calling, collision avoidance, long-range identification and tracking (LRIT), etc. The typical systems are the global maritime distress and safety system (GMDSS), the global navigation satellite systems, and the marine satellite communication systems.

- (1) GMDSS

The global maritime distress and safety system (GMDSS), described in the Safety of Life at Sea (SOLAS) Chapter IV Convention, is an immense integrated global communication network mainly composed of three parts: satellite communication systems (INMARSAT (International Maritime Satellite Organization) and COSPAS-SARSAT (Global Satellite Search and Rescue System)), terrestrial radio communication systems (i.e., coastal radio station), and maritime safety information broadcasting systems. Figure 1 shows its concept map. The system is intended to perform the following functions: alerting (including position determination of the unit in distress), search and rescue coordination, locating, maritime safety information broadcasts, general communications, and bridge-to-bridge communications. Once distress alerts appear, onshore search and rescue agencies and nearby shipping departments will promptly issue alerts through satellite communications and ground communications technology so that they can assist

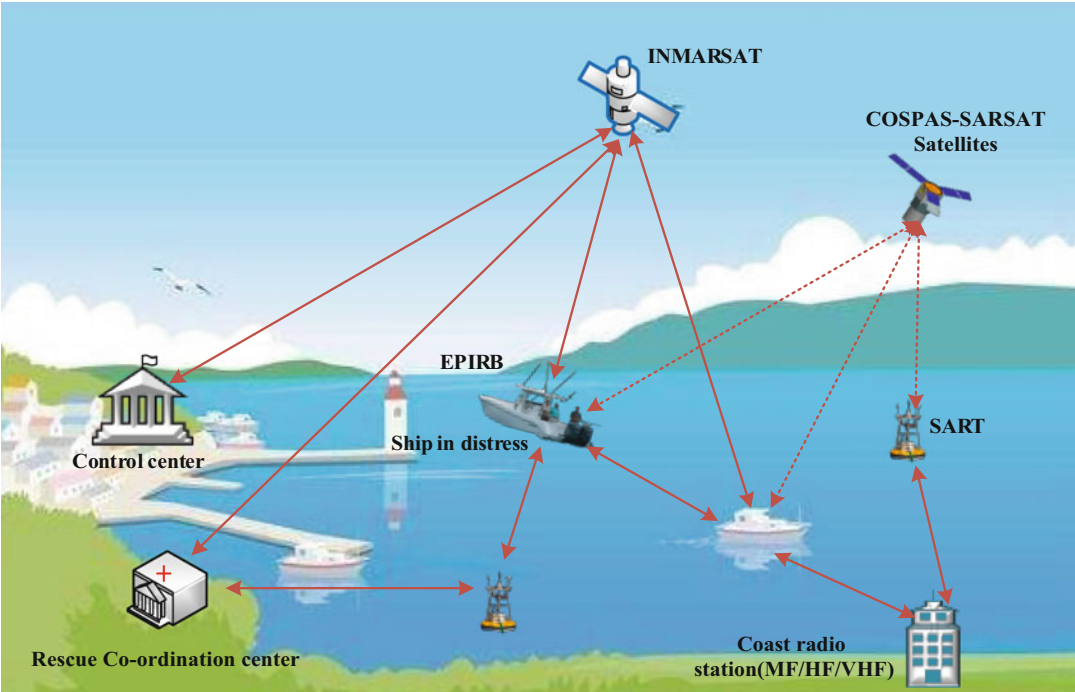
in coordinating search and rescue operations. According to the GMDSS, all cargo or passenger ships with a total tonnage of 300 or more must be equipped with radio equipment that conforms to international standards which guarantees maritime safety to a certain extent.

- (2) The global navigation satellite systems

The global navigation satellite systems (GNSS) is a space-based radio navigation and positioning system that uses satellites to provide autonomous geo-spatial positioning on the earth's surface or near-Earth space. It can be used for providing position and navigation or for tracking the position of something fitted with a receiver (satellite tracking) which is widely applied for the safety of maritime navigation. Common systems include the United States' Global Positioning System (GPS), Russia's GLONASS, China's BeiDou Navigation Satellite System (BDS), and the European Union's GALILEO. Table 1 shows the details.

- (3) The marine satellite communications

The marine satellite communications in the UHF band have global or regional coverage and are commonly deployed on ships and vessels to fulfill different communication requirements despite their high cost. As the commercially provided services, they can implement multiple communication services (voice, data, e-mail, SMS, crew calling, telex, facsimile, remote monitoring, tracking (position reporting), etc.) to guarantee the normal operation of daily cruise and satisfy the needs of crew members to contact with their families. According to the status of satellites, they can be geostationary or non-geostationary. Worldwide, typical marine satellite communication systems mainly include VSAT, Inmarsat systems,



Maritime Internet of Vessels, Fig. 1 Concept map of GMDSS

Maritime Internet of Vessels, Table 1 The global navigation satellite systems

System	BDS	GPS	GLONASS	GALILEO
Coding	CDMA	CDMA	FDMA	CDMA
Frequency (GHz)	1.561098 (B1) 1.589742 (B1–2) 1.20714 (B2) 1.26852 (B3)	1.563–1.587 (L1) 1.215–1.2396 (L2) 1.164–1.189 (L5)	1.593–1.610 (G1) 1.237–1.254 (G2) 1.189–1.214 (G3)	1.559–1.592 (E1) 1.164–1.215 (E5a/b) 1.260–1.300 (E6)
Precision	10 m (public) 0.1 m (encrypted)	15 m (no DGPS or WAAS)	4.5–7.4 m	1 m (public) 0.01 m (encrypted)

iridium satellite system, etc. Table 2 mainly summarizes the correlative systems widely employed in maritime communication.

In addition, it should be noted that as a supplement to conventional communications and emergency communications, coastal mobile communications mixing with cellular or other wireless communication technologies have unique communication advantages in offshore areas. The off-shore coverage of these shore-based communication systems could provide reliable communication guarantee for ports, wharfs, channel manage-

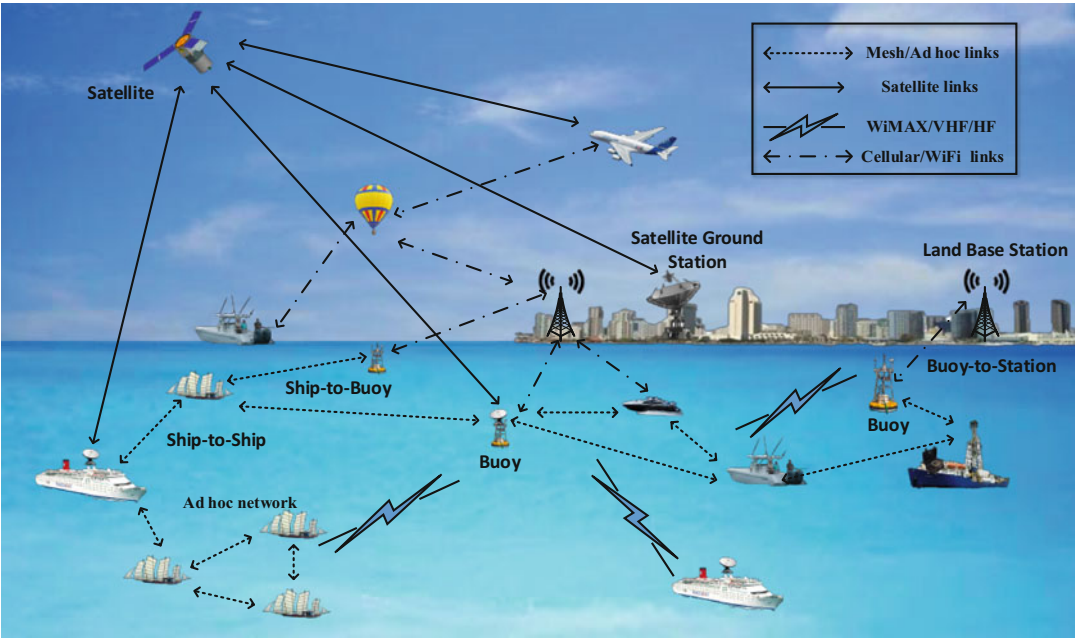
ment, mariculture, and salvage at sea (Minghua et al. 2017), such as GSM/GPRS, 3G, 4G, Wi-Fi, WiMAX, mesh networks, ad hoc networks, and short-range devices like ZigBee and Bluetooth links. Although they could offer higher transmission rate, the communication distance is limited in range. A heterogeneous network integrated with above communication technologies is shown in Fig. 2.

Further Development

Currently, with the high development of maritime industry and activities, the demand for maritime communications services is increasing, which

Maritime Internet of Vessels, Table 2 The marine satellite communication systems

Systems	Services	Data rate	Communication distance
EPIRB	Alarm, identification, positioning, and locating	400 bps	Long distance
Iridium	Positioning, voice, data, paging, short messages, Internet services	Data transmission 2.4 kbps; Internet services 9.6 kbps	Long distance
VSAT	Data, voice, image, video, Internet services	Downlink 64 MbpsUplink 2 Mbps	
Inmarsat	Voice, fax, data, video, etc.	9.6~128 kbps	
Global Xpress(Inmarsat-5)	Voice, data, video, VOIP, etc.	Downlink 50 MbpsUplink 5 Mbps	
Globalstar	Voice, fax, data, short information, location, etc.	Voice 2.4/4.8/9.6 kbpsData 7.2 kbps	



Maritime Internet of Vessels, Fig. 2 A heterogeneous network for Maritime Internet of Vessels

puts forward higher requirements for the safety, reliability, and effectiveness of maritime communications. The traditional communication systems can hardly afford for the extreme volume of data generated by navigation services, video monitoring, and multimedia downloading, and new technologies and schemes are in the pipeline. Following the Internet of Things and Internet of Vehicles, the Maritime Internet of Vessels has come into people’s vision. Maritime autonomous surface ships (MASS) are playing

a very important role in the development of Maritime Internet of Vessels, and many new technologies are being explored to enhance the communication efficiency for the Maritime Internet of Vessels.

• MASS

In recent years, ship intellectualization has become the general trend of global intelligent shipping. In order to reduce the difficulty of ship control and management, reduce

man-made misoperation, improve the safety of equipment and ship operation, optimize ship navigation, reduce costs, and increase profits, the research on MASS has been carried out worldwide (<http://www.sumia.org/newsdetails.aspx?id=2760>).

According to the definition of “intelligent ship specification” promulgated by China Classification Society (CCS) in 2016, “intelligent ships” refers to the use of sensors, communications, Internet, and other technical means to automatically perceive and obtain information and data of the ship itself, marine environment, logistics, ports, and other aspects. Based on computer technology, automatic control technology, and large data processing and analysis technology, the intelligent operation of ships in the aspects of navigation, management, maintenance, and cargo transportation should be realized, so as to make the ships safer, more environmentally friendly, more economical, and more reliable. The functions of intelligent ships are composed of intelligent navigation, intelligent hull, intelligent engine room, intelligent energy efficiency management, intelligent cargo management, and intelligent integrated platform (https://blog.csdn.net/weixin_37978606/article/details/80957137). Figure 3 summarized four development phases for MASS. Now it is in the transitional stage from the first phase to the second phase, and the fourth phase is the era of full automation.

In order to realize the full intellectualization of ships, there are still many problems that had to be overcome. First of all, the communication volume is large because of the need to transmit a large number of sensor information and equipment status information, as well as radar images, sea video, and so on. The maritime communication systems should be able to fulfill the requirements of high bandwidth, low delay, low cost, etc. Besides, adequate attention should be paid to the maritime network security. With the increasing of ship-shore communication, more and more ships are exposed which makes the security

of maritime communication systems becomes increasingly important. How to avoid hacker attacks, key information leaks, and emergency plans of MASS in network attacks are all issues that need further study.

- Application of new technologies

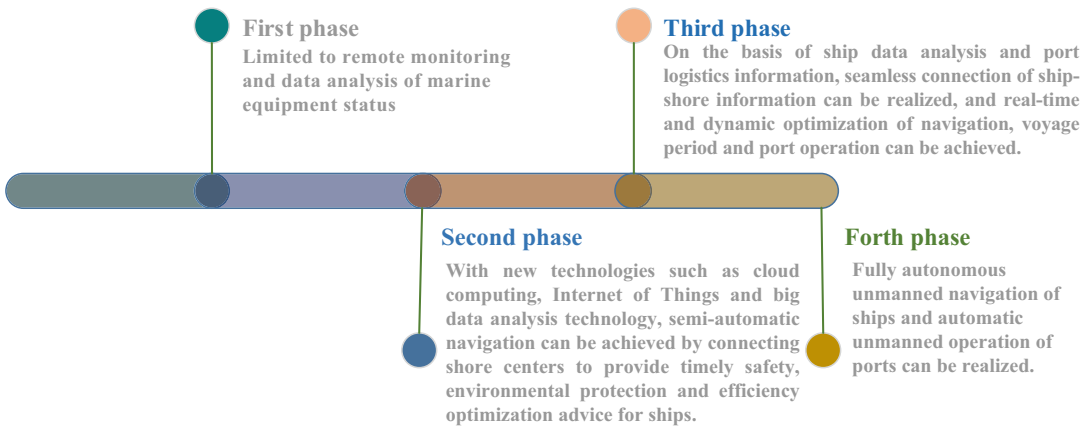
With the refinement, comprehensiveness, and humanization of ship management, the amount of data is increasing rapidly. How to store and manage massive data effectively has become more and more important. The establishment of Maritime Internet of Vessels needs to be data-centric, combined with Internet technologies and Internet of Things technologies, to realize the interconnection between vessels and vessels, people and vessels, vessels and cargo, and vessels and shore-based facilities. For example, the fog/edge computing, information perception and interaction technologies, and big data fusion and analysis technologies can be applied in the field of Maritime Internet of Vessels.

- (1) Fog/edge computing

Currently, continuous development of maritime business has fuelled faster growth in the amount of data and information acquired from ships, cargo, operators, and navigation environment. It has become increasingly important to achieve effective transmission, storage, and management of these massive data although they require massive computing resources, storage space, and communication bandwidth. As an extension of cloud computing, fog/edge computing can be employed to extend computing, storage, and networking resources to the network edge and alleviate the heavy burden of large amounts of data on the network (Ni et al. 2018).

- (2) Information perception and interaction technology

The application of Maritime Internet of Vessels cannot be separated from the support of information sensing technology. With the increasing concentration of maritime traffic routes, the increasing total



Maritime Internet of Vessels, Fig. 3 Development of MASS (<http://www.sumia.org/newsdetails.aspx?id=2760>)

number of vessels, and the sharp rise of communication equipment, information transmission and interaction between different equipment become more and more important, which puts forward higher requirements for information collection, identification, management, and integration. In order to reduce information redundancy and error, improve the information system of ship networking, and improve the intelligent level of Maritime Internet of Vessels, it is necessary to strengthen the research of information perception and interaction technology.

(3) Software-defined network

SDN (software-defined network) is a new network architecture, which can separate network control from traditional hardware devices. Since most of the maritime communication nodes are in motion, the business environment of the network is constantly changing. For the traditional network, if the business needs change, it is quite tedious to modify the configuration of network equipment. Therefore, in order to better meet the special requirements of Maritime Internet of Vessels, the flexibility and agility of the network become more and more important, and SDN can be taken into account to address such issues.

(4) Big data analysis technology

With the continuous development of maritime communications, marine information sensing and communicating equipment, which covers a variety of information transmitting means such as acoustic, optical, and electromagnetic, is continuously generating massive data which have shown the characteristics of huge volume, various data types, fast flow speed, and high value (You and Wei 2018). How to analyze and process these data and improve the breadth and depth of data utilization is a critical issue. It is urgent to strengthen the research of big data analysis technologies, so as to promote the rapid development of Maritime Internet of Vessels.

In addition, the layout of customized systems for Maritime Internet of Vessels should be accelerated, the integration of ship-shore intelligent information should be realized, and a visual sharing platform of information for Maritime Internet of Vessels to ensure the safety of ship navigation and intra-ship communication can also be taken into account. At the same time, the goals of saving energy, reducing operation and maintenance costs, and improving operational efficiency can be achieved. With all kinds of possible network access and interconnection, the ubiquitous links

between vessels and vessels and people can be realized; the intelligent perception, recognition, and management of vessels and processes can also come true, which makes vessels and vessels, vessels and waterways, and vessels and people more interoperable; and the maritime navigation will become more intelligent, controllable, safe, and environmentally friendly.

Conclusion

With the development of shipping industry, the amount of data and information acquired from vessels, cargo, operators, and navigation environment is increasing, and the traditional maritime communication systems can hardly fulfill the development requirements at present. This entry makes a detailed investigation about the existing networking modes for Maritime Internet of Vessels and points out the further development. In 5G era, great efforts should be made to connect more information units for Maritime Internet of Vessels and provide more information sharing services through real-time interconnection, and then the development of intelligent shipping can be further promoted, so that the safety and efficiency of shipping become predictable.

Cross-References

- [Edge Data Centers in Fog Computings](#)
- [Maritime Internet of Vessels](#)
- [MIMO Satellite](#)
- [Software-Defined Wireless Networking](#)

References

- de Sousa MM, Pelas EG (2011) Digital selective calling system for controlling of communications by terrestrial waves on the Maritime mobile service. In: 2011 SBMO/IEEE MTT-S international microwave and optoelectronics conference (IMOC 2011), Natal, pp 612–617
- Lázaro F, Raulefs R, Wei W et al (2017) VHF data exchange system (VDES): an enabling technology for maritime communications. *Ceas Space J* 2017(4): 1–9
- Minghua X, Youmin Z, Erhu C et al (2017) Current status and challenges of marine communications. *Chin Sci Inf Sci* 2017(06):5–23
- Ni J, Zhang K, Lin X, Shen XS (2018) Securing fog computing for internet of things applications: challenges and solutions. *IEEE Commun Surv Tutor* 20(1):601–628
- Weintrit A, Weinrit A (2011) Development of the IMO e-Navigation concept – common Maritime data structure[J]. *Commun Comput Inf Sci* 239:151–163
- You H, Wei Z (2018) Big data technology for maritime information sensing. *Commun Inf Syst Technol* 9(2):1–7

Market Entry

Xuehe Wang and Lingjie Duan
Engineering Systems and Design Pillar,
Singapore University of Technology and Design,
Singapore, Singapore

Synonyms

[Market entry strategy](#)

Definitions

Market entry is the activities associated with bringing a product or service to a targeted market. A market can denote the *primary market* and *secondary market*. The *primary market* refers to the market where the companies sell new products to the customers for the first time, while the *secondary market* is one in which the customers trade the products among themselves.

Historical Background

As the rapid development of the wireless technology, numerous wireless-related services (e.g., 4G wireless service, femtocell service, mobile data trading service) are emerging. When facing the new services, asymmetric behaviors of the wireless service providers (WSPs) and users are observed. For example, the providers may have different timing of introducing the new wireless

services to the market, and the heterogeneous users have different response to these services. Game theory provides solid mathematic tools for analyzing the behaviors of providers and users in wireless networks (Han 2012). One of the most popular examples of game theory in wireless networks is the power control problem, where the transmit power level of a wireless user can pose a positive or negative impact on the transmission rate and quality of service (QoS) of the other users due to interference. By noting the competition between the users, Shah et al. (1998) first formulate the interactions of wireless users as a non-cooperative game and obtain the equilibrium transmit power of all users. Following the pioneer work on non-cooperative games in power control, game theory has been implemented to plenty of application areas in wireless network market entry, e.g., the adoption of femtocell service (Shetty et al. 2009; Yun et al. 2011), new wireless technology introduction (Sen et al. 2010), and secondary mobile data trading market (Zheng et al. 2015; Wang et al. 2016). Musacchio et al. (2006) develop a game theoretic model for studying the upgrade timing of two interconnected Internet service providers (ISPs), where the network upgrade of one ISP has positive network effect on the other. This “free-rider” effect makes it reluctant for the providers to upgrade and instead enjoy the benefits of other providers’ upgrading. Duan et al. (2015) further discuss the upgrade timing of 4G network by taking the user churn into consideration, which weakens the free-riding effect. When introducing new technologies to the network, the competition between entrant and incumbent network technologies may exist. Dynamic games can be used to analyze the users’ adoption dynamics of a new network technology in the presence of an incumbent technology (Sen et al. 2010). To understand the interactions between the providers and their users, Stackelberg games provide an ideal tool to study the hierarchical decision-making among the players, in which the providers (leaders) declare and announce their strategies first and then the users (followers) react selfishly based on the strategies of the providers. Using non-cooperative Stackelberg games, Duan et al. (2013, 2016) study the

economic incentive for an operator to introduce the femtocell service on top of its existing macrocell service centrally and distributedly, in which users submit their bandwidth demands to the operator based on the service and corresponding price and the macrocell operator and femtocell operator (same in central case) can adjust the bandwidth price to maximize their profits. Wang et al. (2016) also use Stackelberg games to study a secondary market in which mobile data users are allowed to trade unused data.

Foundations

Game theory is a basic theory in economics used for the analysis of strategic decision-making of intelligent rational players. The normal-form representation of a game consists of three parts (Gibbons 1992):

- The game players $\mathcal{N} = \{1, 2, \dots, N\}$.
- The available strategy set S_i of each player $i \in \mathcal{N}$. Denote the strategy chosen by player i as $s_i \in S_i$ and the strategies of all players except player i as s_{-i} . Let $S = \prod_i S_i$ denote the set of all possible strategy profiles.
- The utility U_i received by player i for each possible combination of strategies $s \in S$.

Non-cooperative Games

There are many types of games possessing different properties suited for analyzing variety of situations. *Non-cooperative games* are used ubiquitously for the analysis of market entry in wireless networks, in which many players compete for finite resources and each player behaves selfishly and independently by maximizing his own utility, given the possible strategies of other players and their impact on the player’s utility (Osborne 2004). For example, the wireless providers compete over new wireless technologies and pricing strategies to attract users, and the users are involved in many non-cooperative situations such as allocation of bandwidth, transmit power, etc.

To describe the best strategy for a typical player under any given strategy profile of all

other players, the concept of *best response* is introduced.

Definition 1 For each player i , the best response is the strategy such that

$$BR_i(s_{-i}) := \{s_i \in S_i : U_i(s_i, s_{-i}) \geq U_i(s'_i, s_{-i}), \forall s'_i \in S_i\} \quad (1)$$

Nash equilibrium is the stable outcome of a game, which is defined as the strategy profile of all players where none of the players can improve its utility by a unilateral move.

Definition 2 A profile $s^* \in S$ is called a pure strategy Nash equilibrium, if for each player $i \in \mathcal{N}$:

$$U_i(s_i^*, s_{-i}^*) \geq U_i(s_i, s_{-i}^*), \quad \forall s_i \in S_i. \quad (2)$$

Stackelberg Games

In many non-cooperative games, a hierarchy among the players might exist where a fraction of the players (leaders) make decisions first and the rest of the players (followers) react selfishly based on the strategies of the leaders. This kind of games forms special classes of non-cooperative games called *Stackelberg games*, which is widely used in economic literature (He et al. 2007).

For a two-stage Stackelberg game, let $\mathcal{N}_L$ denote the set of leaders who announce their strategies first at Stage I and $\mathcal{N}_F$ denote the set of followers who make decisions accordingly at Stage II. Let $s_{i,L}$ denote the strategy of leader $i \in \mathcal{N}_L$ and s_L denote the strategy profile of all leaders. Let $s_{i,F}$ denote the strategy of follower $i \in \mathcal{N}_F$ and s_F denote the strategy profile of all followers. Then, the concept of Stackelberg equilibrium can be characterized as follows.

Definition 3 A profile $\{s_L^*, s_F^*\} \in S$ is a Stackelberg equilibrium strategy, if for each follower $i \in \mathcal{N}_F$:

$$U_i(s_{i,F}^*, s_{-i,F}^*, s_L^*) \geq U_i(s'_{i,F}, s_{-i,F}^*, s_L^*), \quad \forall s'_{i,F} \in S_i, \quad (3)$$

and for each leader $i \in \mathcal{N}_L$:

$$\begin{aligned} &U_i(s_{i,L}^*, s_{-i,L}^*, s_F^*(s_{i,L}^*, s_{-i,L}^*)) \\ &\geq U_i(s'_{i,L}, s_{-i,L}^*, s_F^*(s'_{i,L}, s_{-i,L}^*)), \\ &\forall s'_{i,L} \in S_i. \end{aligned} \quad (4)$$

Generally, the non-cooperative Stackelberg game is used to analyze the strategic interaction between the wireless providers (who decide the market entry timing and pricing of the services) and wireless users (who decide whether to subscribe to the new services and how much to buy based on their utilities) in the wireless market.

Key Applications

Network economics and pricing can be used to model and analyze many market entry problems in wireless networks. Some key applications are summarized in the following.

Market Entry Timing of New Technologies

As with the emergence of new technologies in wireless networks, the wireless providers should decide whether to deploy the new technology to improve the QoS and the upgrade timing. Generally, the cost of new technology upgrade decreases over time as the technology matures. An existing user can switch to the new technology service of the same provider or a different provider according to the switching cost. Being the first to upgrade to new technology, a provider increases its market share but spends more money on the upgrade. Therefore, the provider decides its upgrade time by trading off increased market share and upgrade cost.

Duan et al. (2015) study the upgrade timing of the fourth generation (4G) technology on the basis of third generation (3G) of cellular wireless networks by developing a game theoretic model for analyzing the competitive providers' interactions. The main results show that:

- When the upgrade cost is low, the 4G monopolist upgrades at the earliest available time; otherwise it postpones its upgrade.
- Under no inter-network switching case, i.e., no user switches operators due to high switching cost, competitive operators upgrade simultaneously, no matter how much the upgrade cost is.
- Under practical inter-network switching case, competitive operators select different upgrade times to avoid severe competition. By upgrading early, an operator captures a large market share and the 4G's QoS improvement which can compensate for the large upgrade cost. The other operator, however, postpones its upgrade to avoid severe competition and benefits from cost reduction. The availability of 4G upgrade may decrease both operators' profits because of the increased competition, and paradoxically, their profits may increase with the upgrade cost.

Secondary Service Entry

To improve the QoS and attract more users, the WSPs are eager to bring new technologies and services to the wireless market on top of the existing services, e.g., the introduction of femtocell service on top of the existing macrocell service. Femtocell technology is introduced to solve the poor signal reception problem for indoor users, as the high-frequency and low-power wireless signals from macrocell base stations often have difficulty in effectively traveling through walls. By bringing the femtocell service to the market, the macrocell operator should consider two main questions:

- Can the macrocell operator benefit from the femtocell service?
- How should the macrocell operator allocate and price the spectrum bands?

Users experience different channel conditions and spectrum efficiencies with the macrocell service, but all of them achieve a high spectrum efficiency with the femtocell service

by deploying a femtocell at home. Thus, users have different preferences between macrocells and femtocells and should decide which service to choose and how much bandwidth to request based on the price of each service.

The macrocell operator can provide the femtocell service itself or lease spectrum to a femtocell operator. For the situation when femtocells are managed by the same operator that controls the macrocells, Duan et al. (2013) model the interactions between the operator and users as a two-stage Stackelberg game. In Stage I, the operator determines bandwidth allocated to femtocell and macrocell service and the corresponding prices. In Stage II, each user decides which service to choose and how much bandwidth to purchase. It reveals that:

- If femtocell service has the same maximum coverage as macrocell service, the operator will choose to provide femtocell service only, as this leads to a better user quality of service and a higher operator profit.
- If all users have reservation payoffs which are what they can achieve with the original macrocell service, then the operator will always continue providing the macrocell service (with or without the femtocell service) to ensure that all users achieve payoffs no worse than their reservation payoffs.
- When multiple femtocells can reuse the same spectrum resource, the operator has more incentives to allocate spectrum to femtocell, and the femtocell price will be reduced. Moreover, when femtocell service incurs operational cost to the operator or has a smaller coverage than the macrocell service, the operator will always serve users by dual services. Furthermore, as cost decreases or coverage increases, more users are served by the femtocell service.

For the situation when the macrocell operator leases spectrum to independent femtocell operators, Duan et al. (2016) model the interactions between a macrocell operator, a femtocell operator, and end users as a three-stage dynamic game.

In Stage I, the macrocell operator decides bandwidth allocations to femtocell service and macrocell service and the macrocell price. In Stage II, the femtocell operator decides how much bandwidth to lease from the macrocell operator and the femtocell retail price. In Stage III, each user decides which service to choose and how much bandwidth to request. It reveals that:

- The macrocell operator has more incentive to lease spectrum to the femtocell operator when its spectrum capacity is small, as femtocell services can help cover more users and improve the utilization efficiency of the limited spectrum resource. Otherwise, the macrocell operator has less incentive to enable femtocell service due to the significant competition.
- In general, femtocell service can increase both the total consumer surplus and social welfare. However, some users who originally receive good service in macrocell might experience a payoff drop after the femtocell service is introduced, due to fewer resources being allocated to the macrocell service and a higher macrocell price.

Secondary Data Trading Market Entry

Another key application is the secondary data market entry for mobile data trading. In reality, a lot of data users subscribe to the data plans with certain amount of data quota which will expire in a billing cycle (e.g., 1 month). Some users can easily use up their data quota and may pay for costly data over-usage, while some users cannot use up their data quota and have leftover data. Motivated by users' diverse usage behavior (more or less than the subscribed data quotas), the secondary data trading market appears in which cellular users can trade data plans with each other, e.g., 2CM (2nd exchange market) data trading platform introduced by China Mobile Hong Kong. Yu et al. (2015), Zheng et al. (2015), and Yu et al. (2017a) discuss such a market, where data users submit bids to buy and sell data. The provider serves as the middleman to match buyers and sellers, as well as charges administration fees

and/or pockets the differences between the buyer and seller prices. Moreover, Yu et al. (2017b) study the mobile data trading problem under future data demand uncertainty. Taking the latest market and usage information into consideration, an algorithm is designed to help estimate the users' risk preference and provide trading recommendations dynamically. Wang et al. (2016) further propose a new type of user-initiated network for cellular users to freely trade data plans by leveraging personal hotspots (PHs) without the wireless provider's involvement. A user with data surplus can set up a PH and share the cellular data connection to another user with data deficit in the vicinity. Moreover, the WSP's countermeasures to the PH market and the competition between WSPs are investigated. Following this work, Wang et al. (2017) study the PH-enabled data-plan sharing between local users and travelers by taking into account the information uncertainty at the traveler side. Lacking the selfish PHs' information, the expected cost of the traveler is higher than that under the complete information, but the gap diminishes as the PHs' spatial density increases.

Future Directions

The economics of market entry can be applied to analyze the upgrading time and interactions between the WSPs and users for many new technologies and services, e.g., the introduction of 5G, smart home.

Cross-References

- [Behavioral Economics](#)
- [Dynamic Pricing](#)
- [Game Theory](#)

References

- Duan L, Huang J, Shou B (2013) Economics of femtocell service provision. *IEEE Trans Mobile Comput* 12(11):2261–2273

- Duan L, Huang J, Walrand J (2015) Economic analysis of 4G upgrade timing. *IEEE Trans Mobile Comput* 14(5):975–989
- Duan L, Shou B, Huang J (2016) Capacity allocation and pricing strategies for new wireless services. *Prod Oper Manag* 25(5):866–882
- Gibbons R (1992) A primer in game theory. Harvester Wheatsheaf, New York
- Han Z (2012) Game theory in wireless and communication networks: theory, models, and applications. Cambridge University Press, Cambridge
- He X, Prasad A, Sethi SP, Gutierrez GJ (2007) A survey of stackelberg differential game models in supply and marketing channels. *J Syst Sci Syst Eng* 16(4):385–413
- Musacchio J, Walrand J, Wu S (2006) A game theoretic model for network upgrade decisions. In: *Proceedings of the 44th annual allerton conference on communication, control, and computing*, Monticello, pp 191–200
- Osborne MJ (2004) An introduction to game theory, vol 3. Oxford university Press, New York
- Sen S, Jin Y, Guérin R, Hosanagar K (2010) Modeling the dynamics of network technology adoption and the role of converters. *IEEE/ACM Trans Netw (TON)* 18(6):1793–1805
- Shah V, Mandayam NB, Goodman DJ (1998) Power control for wireless data based on utility and pricing. In: *Proceedings of the ninth IEEE international symposium on personal, indoor and mobile radio communications*, vol 3. IEEE, pp 1427–1432
- Shetty N, Parekh S, Walrand J (2009) Economics of femtocells. In: *Proceedings of IEEE global telecommunications conference (GLOBECOM)*. IEEE, pp 1–6
- Wang X, Duan L, Zhang R (2016) User-initiated data plan trading via a personal hotspot market. *IEEE Trans Wirel Commun* 15(11):7885–7898
- Wang F, Duan L, Niu J (2017) Optimal pricing of user-initiated data-plan sharing in a roaming market. In: *Proceedings of IEEE global telecommunications conference (GLOBECOM)*. IEEE, pp 1–6
- Yu J, Cheung MH, Huang J, Poor HV (2015) Mobile data trading: a behavioral economics perspective. In: *13th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WiOpt)*. IEEE, pp 363–370
- Yu J, Cheung MH, Huang J (2017a) Economics of mobile data trading market. In: *15th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WiOpt)*. IEEE, pp 1–8
- Yu J, Cheung MH, Huang J, Poor HV (2017b) Mobile data trading: behavioral economics analysis and algorithm design. *IEEE J Sel Areas Commun* 35(4):994–1005
- Yun S, Yi Y, Cho DH, Mo J (2011) Open or close: on the sharing of femtocells. In: *Proceedings of IEEE conference on computer communications (INFOCOM)*. IEEE, pp 116–120
- Zheng L, Joe-Wong C, Tan CW, Ha S, Chiang M (2015) Secondary markets for mobile data: feasibility and benefits of traded data plans. In: *Proceedings of IEEE conference on computer communications (INFOCOM)*. IEEE, pp 1580–1588

Market Entry Strategy

► Market Entry

Massive MIMO

Erik G. Larsson and Emil Björnson
Department of Electrical Engineering (ISY),
Linköping University, Linköping, Sweden

Synonyms

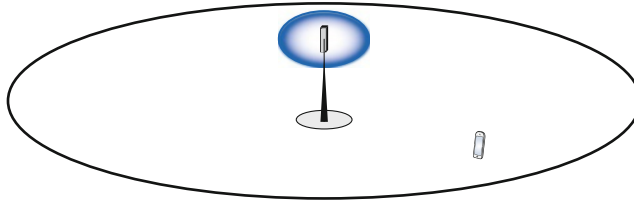
Full-dimension MIMO; Large-scale antenna systems (LSAS); Large-scale MIMO; Massive multiuser MIMO; Very large MIMO

Definition

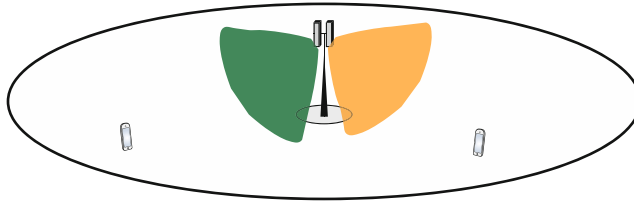
Technology that uses large numbers of phase-coherently operating antennas at wireless base stations to serve a multitude of users by spatial multiplexing.

Historical Background

The wireless transmission from a single antenna has a predefined directivity, independent of where the users is located; see Fig. 1. The idea of using multiple antennas at wireless transceivers goes back a long time, at least to Alexanderson (1919), where analog transmit beamforming was used to direct the signal toward the receiving user and Peterson et al. (1931) where spatial diversity from receive beamforming was discussed. Modern incarnations of the idea have particularly included the concept of spatial-division multiple access (SDMA), proposed in the early 1990s by Swales et al. (1990), Anderson et al. (1991), and Roy and Ottersten (1991), and techniques for the suppression of interference using antenna arrays by Winters (1984). The novelty of these methods was to enable users to be multiplexed spatially,

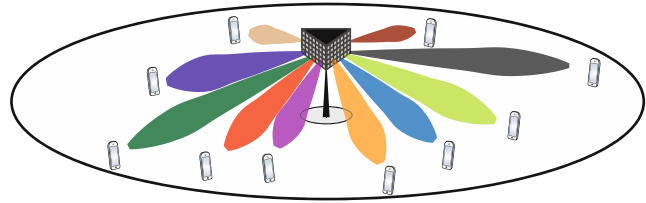


Massive MIMO, Fig. 1 In single-antenna transmission, the transmission has a predefined directivity, independent of the location of the receiving user. One example is omnidirectional transmission, as shown in the figure



Massive MIMO, Fig. 2 In multiuser MIMO transmission with few antennas, the spatially multiplexed signals are broadly directed with beamforming toward the respective receiving users

Massive MIMO, Fig. 3 In Massive MIMO transmission, many users are spatially multiplexed with narrow beamforming, using an array of very many antennas



instead of in time or frequency, to increase the spectral efficiency of a wireless network; see Fig. 2.

Early attempts to efficiently utilize multiple transceiver antennas did not succeed well. There are many potential reasons for this: first, most methods made simplifying assumptions on the propagation environment that do not hold in practice; second, the hardware was not sufficiently advanced at the time; and, third, the corresponding information theory was not yet developed and hence important insights into the significance and difficulties in acquiring accurate channel state information were largely missing. An information-theoretic framework first took form in Caire and Shamai (2003), Goldsmith et al. (2003), and Viswanath and Tse (2003). Subsequently, SDMA became more commonly known as multiuser multiple-input multiple-output (MIMO), which reflects the fact that the

base stations have multiple antennas and there are multiple users.

Despite the mentioned difficulties, elementary forms of SDMA technology were successfully showcased in the mid-1990s by Anderson et al. (1996). ArrayComm in cooperation Kyocera made the first commercial deployment of SDMA as an overlay to the Personal Handyphone System (PHS) in a limited part of South-east Asia.

Foundations

Massive MIMO is an evolved and practically useful form of multiuser MIMO. The key concept is to use a very large number of phase-coherently operating antennas at the base station, as illustrated in Fig. 3. Since its inception in Marzetta (2010), it has quickly evolved from

a wild theoretical concept with an “unlimited” number of antennas to a practical technology that has been demonstrated in field trials, as described in Harris et al. (2017) and Malkowsky et al. (2017) among others and included as one of the main features of the 5G New Radio; see Parkvall et al. (2017).

Exact definitions of “Massive MIMO” differ slightly among different researchers, but the underlying idea is always to equip base stations with arrays of many electronically steerable antennas. These antennas serve many users simultaneously, in the same time-frequency resource. A minimum of 64 base station antennas is commonly considered, and the goal is to support spatial multiplexing of tens of users, each equipped with one or multiple antennas. The very large number of antennas is a design choice to limit inter-user interference and provide a beamforming gain wherever the user is. The textbooks by Marzetta et al. (2016) and Björnson et al. (2017) advocate the use of time-division duplexing (TDD) operation, which permits exploitation of the uplink-downlink reciprocity of the radio propagation. Specifically, the base station uses channel estimates obtained from uplink pilots transmitted by the users to learn the channels in both directions. These estimates are then used for multiuser precoding in the downlink and multiuser detection in the uplink. With TDD operation, Massive MIMO is entirely scalable with respect to the number of base station antennas, in the sense that the amount of resources spent on channel estimation is independent of the number of antennas.

Less restrictive definitions of Massive MIMO are also common. While the preferred operation of Massive MIMO is in TDD mode, where reciprocity between the uplink and downlink holds, operation in frequency-division duplex (FDD) mode is possible in some cases, e.g., using the concepts provided by Adhikary et al. (2013) and Choi et al. (2014), but Flordelis et al. (2018) show that there is a significant performance penalty. While the propagation models in much Massive MIMO literature suggest sub-6 GHz operation, see e.g., Marzetta et al. (2016), Björnson et al.

(2017), and references therein variants of Massive MIMO are also considered for millimeter wave frequency bands.

Three key features of Massive MIMO are as follows:

- High spectral efficiency (measured in bits/s/Hz), primarily attributed to the spatial multiplexing gain
- Provision of uniformly good quality of service in the entire coverage area, by means of beamforming and power control
- Support for high user mobility by efficient channel acquisition (primarily in TDD operation)

The word “massive” refers to the number of antennas and not to the physical size of a base station. Since low-gain antennas are typically used, the antenna arrays can have attractive form factors, e.g., in the 2 GHz band, a half-wavelength-spaced rectangular array with 200 dual-polarized elements is on the order of 1.5×0.75 meters large. At higher carrier frequencies, the arrays will be even smaller, since the dimensions scale with the carrier wavelength. No specific array geometry is required or assumed in Massive MIMO, and distributed antenna deployments are feasible.

The massive number of antennas can give rise to two desirable propagation phenomena. The first is channel hardening, which is a form of massive spatial diversity implying that the link performance fluctuates very little over time and frequency, even when communicating over a fading channel. The second is favorable propagation, which is the ability for the base station to separate the users in the spatial domain; technically, the channel responses associated with different users become nearly orthogonal with an increasing number of antennas.

Information theory, signal processing, and power control for Massive MIMO are well understood, as demonstrated by Marzetta et al. (2016) and Björnson et al. (2017). An important insight is that the link performance over a fading channel can be rigorously quantified in terms of an effective signal-to-interference-and-noise

ratio (SINR), which depends only on the average channel properties between the users and the base stations and on certain system parameters. These expressions take into account the effects of all significant physical phenomena: small-scale and large-scale fading, fading correlation, intra- and intercell interference, channel estimation errors, and pilot interference from pilot reuse in the network (also known as pilot contamination). In the moderately large number-of-antennas regime, the closed-form capacity bounds that result from these effective SINR expressions become convenient proxies for the link performance achievable with practical coding and modulation, power control, and resource allocation schemes.

Massive MIMO is not only a groundbreaking technology for wireless communications, which can vastly increase the spectral efficiency and uniformity of the quality-of-service over previous technologies. It is also an elegant and mathematically rigorous approach to teaching of wireless communications. A comprehensive analytical understanding of the key properties that determine wireless communication link and system performance is facilitated by the capacity bound analysis of Marzetta et al. (2016). Such analyses have simply not been possible before as the corresponding information theory was too complicated for any practical use.

Key Applications

Fifth-generation (and beyond) mobile access networks; Wireless communications with autonomous vehicles; Backhauling in small-cell networks

Cross-References

- [Massive MIMO BDMA Transmission](#)
- [Millimeter Wave Massive MIMO](#)
- [Omnidirectional Transmission for Massive MIMO](#)

- [Per-Beam Synchronization for Millimeter-Wave Massive MIMO](#)
- [Pilot Reuse for Massive MIMO](#)

References

- Adhikary A, Nam J, Ahn JY, Caire G (2013) Joint spatial division and multiplexing—the large-scale array regime. *IEEE Trans Inf Theory* 59(10):6441–6463
- Alexanderson EFW (1919) Transoceanic radio communication. Presented at a joint meeting of the American Institute of Electrical Engineers and the Institute of Radio Engineers, New York
- Anderson S, Millnert M, Viberg M, Wahlberg B (1991) An adaptive array for mobile communication systems. *IEEE Trans Veh Technol* 40(1):230–236
- Anderson S, Forssén U, Karlsson J, Witzschel T, Fischer P, Krug A (1996) Ericsson/Mannesmann GSM field-trials with adaptive antennas. In: *Proceedings of IEEE colloquium on advanced TDMA techniques and applications*, London
- Björnson E, Hoydis J, Sanguinetti L (2017) Massive MIMO networks: spectral, energy, and hardware efficiency. *Found Trends® Signal Process* 11(3–4): 154–655. <https://doi.org/10.1561/20000000093>
- Caire G, Shamai S (2003) On the achievable throughput of a multi-antenna Gaussian broadcast channel. *IEEE Trans Inf Theory* 49(7):1691–1706
- Choi J, Love D, Bidigare P (2014) Downlink training techniques for FDD massive MIMO systems: open-loop and closed-loop training with memory. *IEEE J Sel Topics Signal Process* 8(5):802–814
- Flordelis J, Rusek F, Tufvesson F, Larsson EG, Edfors O (2018) Massive MIMO performance-TDD versus FDD: what do measurements say? *IEEE Trans Wirel Commun* 17(4):2247–2261
- Goldsmith A, Jafar SA, Jindal N, Vishwanath S (2003) Capacity limits of MIMO channels. *IEEE J Sel Areas Commun* 21(5):684–702
- Harris P, Malkowsky S, Vieira J, Bengtsson E, Tufvesson F, Hasan WB, Liu L, Beach M, Armour S, Edfors O (2017) Performance characterization of a real-time Massive MIMO system with LOS mobile channels. *IEEE J Sel Areas Commun* 35(6):1244–1253
- Malkowsky S, Vieira J, Liu L, Harris P, Nieman K, Kundargi N, Wong IC, Tufvesson F, Öwall V, Edfors O (2017) The world's first real-time testbed for Massive MIMO: design, implementation, and validation. *IEEE Access* 5:9073–9088
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wirel Commun* 9(11):3590–3600
- Marzetta TL, Larsson EG, Yang H, Ngo HQ (2016) *Fundamentals of massive MIMO*. Cambridge University Press, Cambridge
- Parkvall S, Dahlman E, Furuskär A, Frenne M (2017) NR: the new 5G radio access technology. *IEEE Commun Stand Mag* 1(4):24–30

- Peterson HO, Beverage HH, Moore JB (1931) Diversity telephone receiving system of R.C.A. communications, Inc. *Proc IRE* 19(4):562–584
- Roy R, Ottersten B (1991) Spatial division multiple access wireless communication systems. US Patent No. 5,515,378
- Swales SC, Beach MA, Edwards DJ, McGeehan JP (1990) The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems. *IEEE Trans Veh Technol* 39(1):56–67
- Viswanath P, Tse DNC (2003) Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. *IEEE Trans Inf Theory* 49(8):1912–1921
- Winters JH (1984) Optimum combining in digital mobile radio with cochannel interference. *IEEE J Sel Areas Commun* 2(4):528–539

Massive MIMO BDMA Transmission

Chen Sun, Xiqi Gao, and Li You
National Mobile Communications Research
Laboratory, Southeast University, Nanjing,
China

Synonyms

[Beam division multiple access \(BDMA\) transmission](#); [Massive MIMO](#); [Statistical channel state information](#)

Definitions

Beam division multiple access (BDMA) transmission refers to a wireless multiple access scheme, where a base station employs massive antennas to generate beams and communicates with several users simultaneously. Multiple access is achieved in BDMA by assigning beams to individual users, and beams for different users are orthogonal (nonoverlapping).

Historical Background

In developing next-generation wireless networks, massive multiple-input multiple-output (MIMO)

is a promising technology to deliver high spectral and power efficiencies (Rusek et al. 2013; Wang et al. 2014). In massive MIMO system, a base station (BS) employs a large number of antennas to simultaneously and jointly serve multiple mobile users. Assuming an environment of independent and identically distributed propagation MIMO channels, such a massive MIMO configuration has several attractive features (Marzetta 2010) including: (1) downlink channel vectors for different users become asymptotically orthogonal; (2) independent user beamforming can mitigate intra-cell interferences and uncorrelated noises; and (3) transmit power can be asymptotically low.

The aforementioned benefits of massive MIMO depend heavily on the availability of channel state information at the transmitter (CSIT). In practice, the CSIT availability hinges on time-division duplex (TDD) operation through uplink training and the channel reciprocity assumption. Yet, training overhead scales linearly with the total number of user antennas. When the number of users is large or when each user has multiple antennas, the overhead becomes prohibitively high. In addition, as user mobility increases, channel coherence time becomes relatively short. As a result, it becomes much more difficult to obtain accurate instantaneous CSIT for medium- or high-mobility user applications. Although the propagation channel itself is likely reciprocal in TDD, the uplink/downlink RF hardware chains at both BS and mobile transceivers are often not reciprocal (Choi et al. 2014). These practical limitations pose serious challenges to accurate CSIT acquisition in TDD systems, especially when dealing with high user mobility. For frequency-division duplex (FDD) systems without instantaneous channel reciprocity, the required number of independent pilot symbols for CSIT acquisition scales with the number of BS antennas, as does the CSIT feedback overhead (Jindal 2006). Therefore, it is likely to be more difficult to acquire global real-time CSIT for precoding in FDD.

Instantaneous CSIT acquisition represents a significant bottleneck; it is more desirable to exploit statistical CSIT for massive MIMO

systems. Statistical CSI varies over much larger time scales than instantaneous channel parameters. Moreover, the uplink and downlink statistics are usually reciprocal in both FDD and TDD systems (Barriac and Madhow 2004). Therefore, the statistical channel information can be more easily obtained by exploiting reciprocity, even for terminals equipped with multiple antennas.

Beam division multiple access (BDMA) transmission scheme is an effective technology (Sun et al. 2015), which can approach optimal transmission, when only statistical CSIT is available at the BS. In the beam domain, different beams transmit independent signals, and one beam transmits signal to at most one user. The BDMA transmission consists of the following steps: (1) acquiring channel statistics, (2) scheduling users and beams, (3) transmitting pilot and data in the uplink and downlink. By user and beam scheduling, beams assigned to different users are nonoverlapping (orthogonal). Thus, BDMA transmission decomposes massive MU-MIMO channel links to multiple small-scale SU-MIMO interference channel links. For each link, the number of transmit beams is much smaller than the number of antennas, and the channel state information between beams and users can be obtained via channel training. BDMA transmission can cope with the difficulties of instantaneous CSIT acquisition, reduce the inter-user interference, and increase the spectrum efficiency. Moreover, BDMA transmission can reduce the complexity of transceiver design and can be applied in TDD or FDD systems.

Beam Domain Channel

A single-cell system consists of one BS equipped with M antennas, and K users, each with N antennas, randomly distributed in the cell. For a physical channel model, suppose that there are P physical paths between the BS and users, and the p th path of the k th user has an attenuation of $a_{p,k}$, an angle $\varphi_{p,k}$ with the transmit antenna array, and an angle $\theta_{p,k}$ with the receive antenna array. For a wideband channel, after OFDM operation, the channel frequency response matrix in the ℓ th

subcarrier of the k th user is given by (Tse and Viswanath 2005):

$$\mathbf{H}_{k,\ell}^d = \sum_{p=1}^P a_{p,k} e^{-j2\pi d_{p,k}/\lambda_c} \mathbf{e}_r(\theta_{p,k}) \mathbf{e}_t^H(\varphi_{p,k}) e^{-j2\pi \ell \tau_{p,k}}, \quad (1)$$

where the superscript d means downlink, $d_{p,k}$ is the physical distance between transmit antenna 1 and receive antenna 1 along path p , λ_c is the carrier wavelength, and $\tau_{p,k}$ is the propagation delay associated with the p th path. Moreover, $\mathbf{e}_r(\theta) \in \mathbb{C}^{N \times 1}$ satisfying $\|\mathbf{e}_r(\theta)\|_2 = 1$ is the user antenna array response vector corresponding to the angle of arrival (AoA) θ , and $\mathbf{e}_t(\varphi) \in \mathbb{C}^{M \times 1}$ satisfying $\|\mathbf{e}_t(\varphi)\|_2 = 1$ is the BS antenna array response vector corresponding to the angle of departure (AoD) φ .

Assume that the array response vectors corresponding to distinct AoDs are asymptotically orthogonal with infinite number of antennas at the BS (You et al. 2015), i.e.:

$$\lim_{M \rightarrow \infty} \mathbf{e}_t^H(\varphi) \mathbf{e}_t(\eta) = \delta(\varphi - \eta). \quad (2)$$

At the user side, the angle $\theta_{n,k}$ is the sampling of receive signals. When the response vectors are orthogonal, i.e.:

$$\mathbf{e}_r^H(\theta_{n,k}) \mathbf{e}_r(\theta_{n',k}) = \delta(n - n'), \quad (3)$$

users can separate these orthogonal direction signals perfectly. For uniform linear antenna arrays, uniform sampling of $\sin(\theta_{n,k})$, i.e., $\sin(\theta_{n,k}) = n/N$, is a classical choice for spatial angles (Sayeed 2002).

The channel matrix in (1) can be rewritten as

$$\begin{aligned} \mathbf{H}_{k,\ell}^d &= \sum_{n=1}^N \sum_{m=1}^M \left[\tilde{\mathbf{H}}_{k,\ell}^d \right]_{nm} \mathbf{e}_r(\theta_{n,k}) \mathbf{e}_t^H(\varphi_m) \\ &= \mathbf{U}_k \tilde{\mathbf{H}}_{k,\ell}^d \mathbf{V}^H \end{aligned} \quad (4)$$

where $\mathbf{U}_k = [\mathbf{e}_r(\theta_{1,k}), \mathbf{e}_r(\theta_{2,k}), \dots, \mathbf{e}_r(\theta_{N,k})] \in \mathbb{C}^{N \times N}$ is a unitary matrix and $\mathbf{V} = [\mathbf{e}_t(\varphi_1), \mathbf{e}_t(\varphi_2), \dots, \mathbf{e}_t(\varphi_M)] \in \mathbb{C}^{M \times M}$. As the

number of BS antennas M tends to infinity, $\mathbf{V}$ becomes an asymptotically unitary matrix. We call $\tilde{\mathbf{H}}_{k,\ell}^d$ the *beam domain channel matrix* with one direction of AoD called a beam. Following Sayeed (2002), we have the following sampling approximation:

$$\left[\tilde{\mathbf{H}}_{k,\ell}^d\right]_{nm} \approx \sum_{p \in S_{r,n} \cap S_{t,m}} a_{p,k} e^{-j2\pi d_{p,k}/\lambda_c} e^{-j2\pi \ell \tau_{p,k}} \quad (5)$$

where $S_{r,n}$ is the set of all paths whose receive angles are nearest to the angle $\theta_{n,k}$ and $S_{t,m}$ is the set of paths whose transmit angle is exactly equal to the angle φ_m . In the beam domain, different elements of the channel matrix represent the signals of different transmit and receive angles. Different from the summation of all path signals in the physical domain, the beam domain channel can separate the paths of different angles by different beams.

Define the eigenmode channel coupling matrix as (Gao et al. 2009)

$$\boldsymbol{\Omega}_{k,\ell}^d = \mathbb{E} \left\{ \tilde{\mathbf{H}}_{k,\ell}^d \odot (\tilde{\mathbf{H}}_{k,\ell}^d)^* \right\} \quad (6)$$

whose elements specify the mean amount of energy that is coupled from the m th eigenvector of the BS to the n th eigenvector of users. The eigenmode channel coupling matrix $\boldsymbol{\Omega}_{k,\ell}^d$ is likewise independent of subcarriers. Then, when the BS has access only to the channel coupling matrix, the optimal transmit strategies are the same for all subcarriers, and the subscript ℓ can be omitted. This observation will significantly decrease the overhead of CSIT acquisition.

Beam Division Multiple Access Transmission

As the channel gains of different paths are independent, the beam domain channel matrix $\tilde{\mathbf{H}}_k^d$ (whose subscript ℓ is omitted) has uncorrelated elements. The optimal transmission in the beam domain is that BS transmits independent signals

through different beams and that different set of beams transmit signals to different users, which is called beam division multiple access (BDMA) transmission (Sun et al. 2015, 2017; You et al. 2017). In the BDMA transmission, the beams for different users are nonoverlapping (orthogonal), and one beam transmits at most one user's signal.

Figure 1 illustrates the BDMA downlink and uplink transmission. In BDMA downlink transmission, as shown in Fig. 1a, BS employs different beam sets transmitting different user's signal, and the received signal at user k is

$$\begin{aligned} \mathbf{y}_k^d &= [\tilde{\mathbf{H}}_k^d]^{\mathcal{B}_k} \tilde{\mathbf{x}}_k + \sum_{k' \neq k} [\tilde{\mathbf{H}}_k^d]^{\mathcal{B}_{k'}} \tilde{\mathbf{x}}_{k'} + \mathbf{n}_k \\ &= [\tilde{\mathbf{H}}_k^d]^{\mathcal{B}_k} \tilde{\mathbf{x}}_k + \mathbf{n}_k^d, \end{aligned} \quad (7)$$

where $\mathcal{B}_k$ is the beam set for user k , $\tilde{\mathbf{x}}_k = \mathbf{V}^H \mathbf{x}_k$ is the beam domain transmitted signal, and $\mathbf{n}_k^d$ is the aggregate interference plus noise.

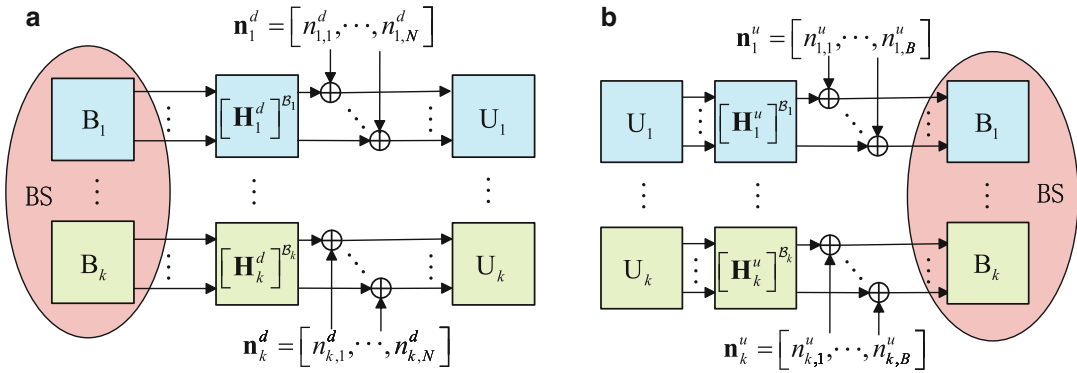
In BDMA uplink transmission, as shown in Fig. 1b, BS employs beam set $\mathcal{B}_k$ to receive signal of user k , and the received signal of user k at the BS is

$$\begin{aligned} \mathbf{y}_k^u &= [\tilde{\mathbf{H}}_k^u]^{\mathcal{B}_k} \tilde{\mathbf{x}}_k + \sum_{k' \neq k} [\tilde{\mathbf{H}}_k^u]^{\mathcal{B}_{k'}} \tilde{\mathbf{x}}_{k'} + \mathbf{n} \\ &= [\tilde{\mathbf{H}}_k^u]^{\mathcal{B}_k} \tilde{\mathbf{x}}_k + \mathbf{n}_k^u, \end{aligned} \quad (8)$$

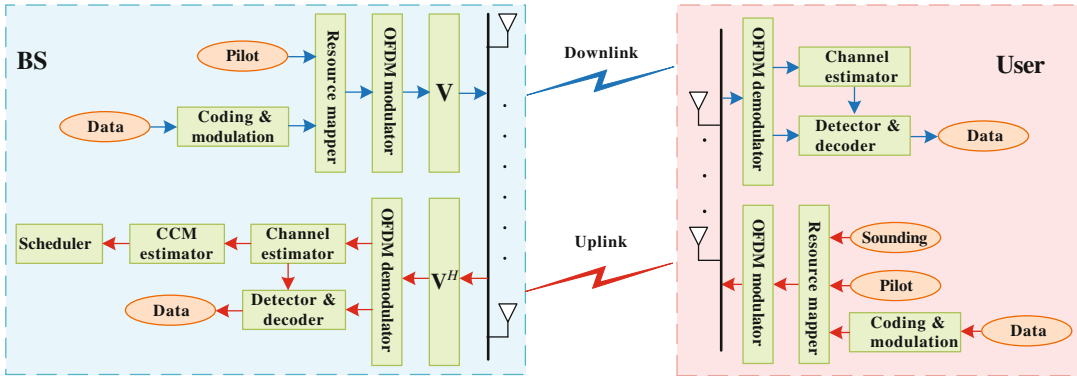
where the superscript u means uplink and $\mathbf{n}_k^u$ is the aggregate interference plus noise of user k in the uplink.

BDMA transmission scheme for MU-MIMO in the beam domain is illustrate in Fig. 2, which consists of the following key components:

1. The BS acquires channel coupling matrices from all users in its own cell. The channel coupling matrices acquisition consists of two steps: first in the uplink, different users transmit sounding signals. Secondly, the BS estimates the instantaneous CSI to calculate channel coupling matrices. Since the statistical CSI varies much slower than the instantaneous CSI, we do not need to estimate the whole channel realizations, and therefore, the



Massive MIMO BDMA Transmission, Fig. 1 (a) BDMA downlink transmission and (b) BDMA uplink transmission



Massive MIMO BDMA Transmission, Fig. 2 Block diagram of a BDMA transmission scheme

overhead of statistical CSI is much less than that of instantaneous CSI.

2. The BS schedules users to be sufficiently separated in beam domain. Based on the channel coupling matrices, the user scheduler selects users for maximizing the sum-rate with the constraints that different user beams are nonoverlapping. After user scheduling, beams at the BS are divided into different sets, leading to decomposing the massive MU-MIMO link into small dimensional SU-MIMO links.
3. Uplink and downlink transmissions consist of pilot training and data transmission. In the training step, since users are separated by different beam sets at the BS, uplink and downlink pilot sequences do not need to be orthogonal. According to the number of selected users, optimal pilot sequences are designed to minimize the channel estimation error. The

receivers in the uplink and downlink estimate the reduced dimensional channel matrices, as well as the aggregate interference covariance matrix for data detection. With channel information, the receivers utilize the iterative soft-input and soft-output (SISO) detection and SISO decoding.

The BDMA scheme is suitable for FDD systems with only statistical CSI at the BS. Note, however, that the approach also applies to TDD systems, when instantaneous CSI at the BS is not available, e.g., for cases involving high user mobility.

Conclusion

BDMA transmission is a multiple access scheme in massive MIMO communication systems. In

the beam domain, the elements of the channel matrix represent the channels between different transmit and receive angles. According to the channel statistics, BS assigns nonoverlapping beams to different users, and users are separated by different beams. Massive MU-MIMO transmission link is decomposed into multiple small dimensional SU-MIMO links, to cope with the difficulties of instantaneous CSIT acquisition and reduce the complexity of transceiver design.

Key Applications

BDMA transmission is an optimal and effective multi-user transmission scheme, when only statistical CSIT is available in both TDD and FDD systems. Employing massive antennas at the BS, BDMA transmission explores the spatial resource and provides high spatial resolution by different beams. BDMA transmission helps to provide high spectral efficiency and reduce the complexity of transceiver design.

Cross-References

- [Massive MIMO](#)
- [Per-Beam Synchronization for Millimeter-wave Massive MIMO](#)
- [Pilot Reuse for Massive MIMO](#)

References

- Barriac G, Madhow U (2004) Space-time communication for OFDM with implicit channel feedback 50(12):3111–3129
- Choi J, Love DJ, Bidigare P (2014) Downlink training techniques for FDD massive MIMO systems: open-loop and closed-loop training with memory. *IEEE J Sel Areas Signal Process* 8(5):802–814
- Gao XQ, Jiang B, Li X, Gershman AB, McKay MR (2009) Statistical eigenmode transmission over jointly correlated MIMO channels. *IEEE Trans Inf Theory* 55(8):3735–3750
- Jindal N (2006) MIMO broadcast channels with finite-rate feedback. *IEEE Trans Inf Theory* 52(11):5045–5060
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wirel Commun* 9(5):3590–3600
- Rusek F, Persson D, Lau BK, Larsson EG, Marzetta TL, Edfors O, Tufvesson F (2013) Scaling up MIMO: opportunities and challenges with very large arrays. *IEEE Signal Process Mag* 30(1):40–60
- Sayed AM (2002) Deconstructing multi-antenna fading channels. *IEEE Trans Signal Process* 50(10):2563–2579
- Sun C, Gao X, Jin S, Matthaiou M, Ding Z, Xiao C (2015) Beam division multiple access transmission for massive MIMO communications. *IEEE Trans Commun* 63(6):2170–2184
- Sun C, Gao X, Ding Z (2017) BDMA in multicell massive MIMO communications: power allocation algorithms. *IEEE Trans Signal Process* 65(11):2962–2974
- Tse D, Viswanath P (2005) Fundamentals of wireless communication. Cambridge University Press, New York
- Wang CX, Haider F, Gao XQ, You XH, Yang Y, Yuan D, Aggoune H, Haas H, Fletcher S, Hepsaydir E (2014) Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Commun Mag* 52(2):122–130
- You L, Gao XQ, Xia X, Ma N, Peng Y (2015) Pilot reuse for massive MIMO transmission over spatially correlated Rayleigh fading channels. *IEEE Trans Wirel Commun* 14(6):3352–3366
- You L, Gao XQ, Li GY, Xia XG, Ma N (2017) BDMA for millimeter-wave/Terahertz massive MIMO transmission with per-beam synchronization. *IEEE J Sel Areas Commun* 35(7):1550–1563

Massive MIMO Channel Estimation

Jie Yang¹, Shi Jin¹, Chao-Kai Wen², and Tao Jiang³

¹National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

²National Sun Yat-Sen University, Kaohsiung, Taiwan

³Huazhong University of Science and Technology, Wuhan, China

Synonyms

[Hyper MIMO channel estimation](#); [Large-scale multiple-antennas channel estimation](#)

Definitions

Massive MIMO channel estimation is a procedure to obtain channel state information (CSI) under systems consisting of large-scale multiple antennas at both transmitter and receiver before detecting the transmitted data. CSI refers to the transmission matrix between the transmitter and receiver, which includes scattering, fading, and power decay effects of a communication link. Massive MIMO channel estimation performance determines the suitability of massive MIMO techniques, offers the possibility of improving link reliability and network capacity, and reduces latency and interference.

Historical Background

Massive MIMO technology has been widely studied and incorporated into many standards, such as long-term evolution (LTE) and IEEE802.11 (Wi-Fi), because it can significantly improve the capacity and reliability of wireless systems Larsson et al. (2014).

Over the past two decades, the speed of wireless networks has improved, starting from SISO in 2000, to SU-MIMO in 2009, and to MU-MIMO in 2012. In the near future, around 2020, massive MIMO aims to achieve a data rate of more than 10 Gbps Andrews et al. (2014). Massive MIMO is a bold vision of development from classical MIMO by using very large-scale antennas at both ends, which are fully coherent and adaptive. Wireless channel diversity is applied to provide links with higher speed and reliability.

The larger the number of antennas placed at both transmitter and receiver, the more terminals can be served by the increasing data streams. This condition is expected to bring significant improvements in throughput. Furthermore, transmitters can easily avoid transmission into undesired directions, thereby alleviating harmful interference. This condition can enhance the interference cancelation and system reliability through spatial diversity Björnson et al. (2016).

The aforementioned advantages and improvements in the massive MIMO technology are based on perfect knowledge of the channel matrix, which has to be estimated. However, the massive MIMO channel estimation faces serious challenges. First, the wireless channel for massive MIMO is highly complicated, being characterized by selectivities in frequency, time, and space Lu et al. (2014). Second, the channel information is acquired by probing finite-length pilot sequences into the wireless channel because with the existence of intercell interference, the reused pilot sequences from neighboring cells would contaminate one another. According to Marzetta T.L. (2010), pilot contamination does not vanish with an unlimited number of antennas. Moreover, the training and feedback overhead become overwhelming with the increase in the number of antennas. These problems emphasize the need for alternative channel estimation schemes for increasingly large transceiver arrays.

The development pathway of the massive MIMO channel estimation should highlight the following popular techniques: To deal with the pilot contamination, Yin et al. (2013) present a novel approach that enables low-rate coordination between cells during the channel estimation phase. This approach can offer a powerful method of discriminating across mutual interfering users with strongly correlated pilot sequences. Furthermore, Yang et al. (2014) propose a new two-way training scheme for discriminatory channel estimation (DCE) by exploiting a whitening-rotation-based semi-blind method, which achieves better DCE performance than the existing DCE schemes based on linear minimum mean square error (LMMSE) channel estimator; this proposed method is robust against pilot contamination attacks. In particular, Wen et al. (2015a) introduce the estimation not only of the channel parameters of the desired links in a target cell but also those of the interference links from adjacent cells. This channel estimation method is based on sparse Bayesian learning methods and utilizes the approximately sparse property of the beamspace channel. In addition, a subspace method for channel estimation, which considers a highly sensible finite-dimensional physical

channel model, has been proposed to address the pilot contamination effect Teeti et al. (2015).

Extensive research has been conducted to exploit compressive sensing (CS) techniques to reduce the training overhead and feedback cost, especially for FDD massive MIMO systems. In Rao et al. (2014), a distributed compressive channel state information at transmitter (CSIT) estimation scheme is presented by exploiting the hidden joint sparsity structure in the user channel matrices. Additionally, a discrete Fourier transform (DFT)-aided spatial basis expansion model is introduced by Xie et al. (2016) to represent the channel, which can be applied to TDD and FDD systems under certain conditions. However, the DFT-based methods are only applicable to uniform linear arrays and have power leakage problems. Subsequently, Dai et al. (2018) design an off-grid model for downlink channel sparse representation with arbitrary 2D-array antenna geometry and apply an efficient sparse Bayesian learning (SBL) approach for sparse channel recovery and off-grid refinement. Also, CS-based massive MIMO channel estimation, such as structured compressive sensing-based spatiotemporal joint channel estimation scheme Gao et al. (2016) and beam-blocked compressive channel estimation scheme Huang et al. (2017), has been effective in reducing the required pilot overhead.

Computational complexity is one of the main challenges for massive MIMO systems. To reduce the computational complexity of the channel estimation, a set of low-complexity Bayesian channel estimators, called polynomial expansion channel estimators, are demonstrated by Shariati et al. (2014). To further reduce the computational complexity and improve the accuracy of channel estimation, Wen et al. (2015b) adopt a joint channel-and-data estimation method based on Bayes optimal inference.

Furthermore, other concepts, such as beam-domain channel estimation Duly et al. (2014); Xiong et al. (2017), eigenvalue-decomposition-based channel estimation Xu et al. (2017), and blind channel estimation Ngo et al. (2015), have also received considerable attention.

Foundations

Massive MIMO channel estimation methods are usually divided into two categories Hassan et al. (2017):

- *Training-based channel estimation*, which transmits known training symbols (also called pilots) at predetermined times and frequencies known by the receiver. Then, the receiver uses the known information to obtain the CSI (including the gain and phase rotation imparted by the channel at each point in time and frequency) based on the characteristics of the received training symbols.
- *Blind-based channel estimation*, which estimates the channel without the assistance of known training symbols. This method can save timefrequency resources and offer higher bandwidth efficiency. As the blind-based channel estimation method has lower speed and poorer performance than the training-based method, researchers pay closer attention to the latter. Thus, the later part of this section considers only the training-based channel estimation methods.

Two common modes for massive MIMO systems exist, namely, time division duplexing (TDD) and frequency division duplexing (FDD) Lu et al. (2014). In FDD-mode massive MIMO systems, the uplink and downlink use different frequency bands. Therefore, the uplink and downlink CSI is different. The uplink channel estimation is conducted at the base station (BS) by enabling all users to send different pilot sequences. The downlink channel estimation is followed by two steps: the BS sends pilot sequences to all users, and then the users feed the estimated downlink CSI back to the BS. As the number of antennas increases, the consumption of time resource required to transmit pilot sequences also increases in the FDD mode, whereas the TDD mode is more efficient in this aspect. In TDD systems, based on the assumption of channel reciprocity, only the uplink channel estimation is needed. According to this protocol,

the BS estimates CSI by using the pilot sequences sent by all users to estimate CSI, and then the BS uses the estimated CSI to detect the uplink data and generate beamforming vectors for downlink data transmission. The time required to transmit pilot sequences is less in the TDD mode than in the FDD mode; thus, the TDD mode is usually assumed in the study.

Pilot design is one of the key steps in massive MIMO channel estimation. Pilot placement has to comply with the following principles: the time interval between pilot symbols should be less than the coherence time of the channel and the frequency spacing should be less than the coherence bandwidth Hampton (2013). These conditions mean that in a fast-fading environment, pilots have to be placed relatively often in time, and in a highly frequency-selective channel, pilots have to be placed close together in the frequency dimension. Furthermore, based on meeting the aforementioned principles, pilots should be spaced as far apart as possible to reduce the training overhead. In massive MIMO systems, the received signal at each antenna is the superposition of the signals from all of the transmission antennas. Thus, the pilot sequences have to maintain temporal, frequency, or signal orthogonality to avoid interfering with one another. A common pilot placement strategy is packet-based in which synchronization and pilot sequences are placed in the header of the packet and then the body of the packet contains the data. Based on the hypothesis that the channel only changes slightly in the duration of a packet, the receiver can estimate the channel at the beginning of a packet and then use that estimation during the body of the packet to detect data.

Massive MIMO channel estimation techniques can be elaborated in terms of sub-6 GHz and millimeter wave (mmWave). The foundational channel estimation techniques for sub-6 GHz are introduced as follows:

- *Narrowband MIMO channel estimation:* For the quasi-static fading channel, where the packet length is shorter than the coherence time, the channel is assumed to be changeless

over one packet. Therefore, the transmitter can transmit an N_t -symbol packet, including N_p pilots, every T_{coh} seconds. Then, the receiver can estimate the channel through pilots and use the estimated channel to detect the data in the same packet. The received data are given by

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}, \quad (1)$$

where $\mathbf{H}$ is the $N_r \times N_t$ channel matrix, $\mathbf{X}$ is the $N_t \times p$ pilot matrix, and $\mathbf{N}$ is the $N_r \times p$ noise matrix. The commonly used estimation techniques are the following:

- **Maximum likelihood (ML)** channel estimation aims to maximize the likelihood function of $\mathbf{H}$, namely, $\hat{\mathbf{H}}_{ML} \triangleq \arg \max_{\{\mathbf{H}\}} p(\mathbf{Y}|\mathbf{H})$. Then, after simplification, the solution of $\hat{\mathbf{H}}_{ML}$ is given by $\hat{\mathbf{H}}_{ML} = (\mathbf{Y}\mathbf{X}^H)(\mathbf{X}\mathbf{X}^H)^{-1}$.
- **Least-square (LS)** channel estimation minimizes the squared error between the actual received signal $\mathbf{Y}$ and the estimated received signal $\hat{\mathbf{Y}} = \hat{\mathbf{H}}\mathbf{X}$, that is, $\hat{\mathbf{H}}_{LS} \triangleq \arg \min_{\{\hat{\mathbf{H}}\}} \|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2$. Then, after simplification, the solution of $\hat{\mathbf{H}}_{LS}$ is expressed by $\hat{\mathbf{H}}_{LS} = (\mathbf{Y}\mathbf{X}^H)(\mathbf{X}\mathbf{X}^H)^{-1}$, which is equivalent to the ML channel estimation.
- **Linear minimum mean square error (LMMSE)** channel estimation minimizes the mean square error between the actual and estimated channels; this error is expressed as $\hat{\mathbf{H}}_{LMMSE} \triangleq \arg \min_{\{\hat{\mathbf{H}}\}} \mathbb{E} \left\{ \|\mathbf{H} - \hat{\mathbf{H}}\|_F^2 \right\}$. After simplification, the solution of $\hat{\mathbf{H}}_{LMMSE}$ is given by $\hat{\mathbf{H}}_{LMMSE} = \mathbf{Y}(\mathbf{I} + \mathbf{X}^H\mathbf{X})^{-1}\mathbf{X}^H$.
- **Maximum a posteriori (MAP)** channel estimation is designed to maximize the posterior probability distribution function, that is, $\hat{\mathbf{H}}_{MAP} \triangleq \arg \max_{\{\mathbf{H}\}} p(\mathbf{H}|\mathbf{Y})$.

The ML and LS channel estimations are unbiased estimation techniques as long as the

noise has a zero-mean symmetrical distribution. The LMMSE channel estimation is not unbiased, whereas with the increase in signal-to-noise ratio, it becomes unbiased and is equivalent to the ML/LS estimation.

- *Broadband MIMO channel estimation:* When the bandwidth of the signal is greater than the channel coherence bandwidth, the frequency-selective fading channel occurs. The orthogonal frequency division multiplexing (OFDM) technique is commonly used in broadband massive MIMO applications. In OFDM systems, the pilot is placed in both time and frequency domains, and the frequency domain channel is usually estimated by ML/LS/LMMSE channel estimation techniques.
- *CS-based mmWave massive MIMO channel estimation:* CS techniques bring numerous benefits to mmWave massive MIMO beamspace channel estimation, by using CS channel estimation algorithms. The system operation requires a relatively small training overhead. To adapt to various channel estimation scenarios, several CS algorithms have been introduced, such as LASSO Tibshirani (1996), orthogonal matching pursuit Cai et al. (2011), support detection Gao et al. (2017), and sparse non-informative parameter estimator-based cospase analysis approximate message passing for imaging Yang et al. (2018) channel estimation algorithms.

Finally, the bit error ratio and normalized mean square error are commonly used to study the performance of massive MIMO channel estimation techniques.

Key Applications

Massive MIMO channel estimation is involved in most wireless communication systems, especially in 5G mobile communications.

Cross-References

- [Massive MIMO](#)
- [Millimeter wave massive MIMO](#)

References

- Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong AC, Zhang JC (2014) What will 5G be? *IEEE J Sel Areas Commun* 32(6):1065–1082
- Björnson E, Larsson EG, Marzetta TL (2016) Massive MIMO: ten myths and one critical question. *IEEE Commun Mag* 54(2):114–123
- Cai TT, Wang L (2011) Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans Inf Theory* 57(7):4680–4688
- Dai J, Liu A, Lau VKN (2018) FDD massive MIMO channel estimation with arbitrary 2D-array geometry. *IEEE Trans Signal Process* 99(99):1–1
- Duly AJ, Kim T, Love DJ, Krogmeier JV (2014) Closed-Loop beam alignment for massive MIMO channel estimation. *IEEE Commun Lett* 18(8):1439–1442
- Fang J, Li X, Li H, Gao F (2017) Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems. *IEEE Trans Wirel Commun* 16(3):1935–1947
- Gao Z, Dai L, Dai W, Shim B, Wang Z (2016) Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO. *IEEE Trans Commun* 64(2):601–617
- Gao X, Dai L, Han S, I C-L, Wang X (2017) Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array. *IEEE Trans Wirel Commun* 16(9):6010–6021
- Hampton JR (2013) Introduction to MIMO communications. Cambridge University Press, Cambridge
- Hassan N, Fernando X (2017) Massive MIMO wireless networks: an overview. *Electronics* 6:63
- Huang W, Huang Y, Xu W, Yang L (2017) Beam-blocked channel estimation for FDD massive MIMO with compressed feedback. *IEEE Access* 5:11791–11804
- Larsson E, Edfors O, Tufvesson F, Marzetta T (2014) Massive MIMO for next generation wireless Systems. *IEEE Commun Mag* 52(2):186–195
- Lu L, Li GY, Swindlehurst AL, Ashikhmin A, Zhang R (2014) An overview of massive MIMO: benefits and challenges. *IEEE J Sel Top Signal Process* 8(5):742–758
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wirel Commun* 9(11):3590–3600

- Ngo HQ, Larsson EG (2015) Blind estimation of effective downlink channel gains in massive MIMO. In: Proc ICASSP 2015, Brisbane, pp 2919–2923
- Rao X, Lau VK (2014) Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO Systems. *IEEE Trans Signal Process* 62(12):3261–3271
- Shariati N, Bjornson E, Bengtsson M, Debbah M (2014) Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics. *IEEE J Sel Top Signal Process* 8(5):815–830
- Teeti M, Sun J, Gesbert D, Liu Y (2015) The impact of physical channel on performance of subspace-based channel estimation in massive MIMO systems. *IEEE Trans Wirel Commun* 14(9):4743–4756
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodol)* 58:267–288
- Wen CK, Jin S, Wong KK, Chen JC, Ting P (2015a) Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning. *IEEE Trans Wirel Commun* 14(3):1356–1368
- Wen CK, Wang CJ, Jin S, Wong KK, Ting P (2015b) Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs. *IEEE Trans Signal Process* 64(10):2541–2556
- Xie H, Gao F, Zhang S, Jin S (2016) A simple DFT-aided spatial basis expansion model and channel estimation strategy for massive MIMO systems. In: Proceedings of GLOBECOM, 2016, Washington, DC, pp 1–6
- Xiong X, Wang X, Gao, X, You X (2017) Beam-domain channel estimation for FDD massive MIMO systems with optimal thresholds. *IEEE Trans Wirel Commun* 16(7):4669–4682
- Xu W, Xiang W, Jia Y, Li Y, Yang Y (2017) Downlink performance of massive-MIMO systems using EVD-based channel estimation. *IEEE Trans Veh Technol* 66(4):3045–3058
- Yang J, Xie S, Zhou X, Yu R, Zhang Y (2014) A semiblind two-way training method for discriminatory channel estimation in MIMO systems. *IEEE Trans Commun* 62(7):2400–2410
- Yang J, Wen CK, Jin S, and Gao F (2018) Beamspace channel estimation in mmwave systems via cospars image reconstruction technique. *IEEE Trans Commun*, to be published, DOI [10.1109/TCOMM.2018.2805359](https://doi.org/10.1109/TCOMM.2018.2805359)
- Yin H, Gesbert D, Filippou M, Liu YA (2013) Coordinated approach to channel estimation in large-scale multiple-antenna systems. *IEEE J Sel Areas Com* 31(2): 264–273

Matching for Cooperative Spectrum Sharing

Lin Gao¹, Lingjie Duan², Shimin Gong³, and Qinyu Zhang¹

¹School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China

²Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore

³Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Synonyms

Cooperative spectrum sharing; Dynamic spectrum access; Two-sided matching

Definition

Dynamic spectrum access, also known as opportunistic spectrum access, is a new spectrum access paradigm, which allows unlicensed secondary users (SUs) to access the licensed spectrum of primary users (PUs) in an opportunistic way, and hence can effectively increase the spectrum efficiency and alleviate the spectrum scarcity. *Cooperative spectrum sharing* is an effective form of dynamic spectrum access, where SUs relay the transmissions of PUs in exchange for the opportunity of accessing the licensed spectrum of PUs, and thus, it can offer the necessary incentives for both PUs and SUs in dynamic spectrum access. *Two-sided matching*, or matching, is an efficient framework for analyzing the inter-dependent interactions between two disjoint player sets such as PUs and SUs in cooperative spectrum sharing.

Massive Multiuser MIMO

► [Massive MIMO](#)

Historical Background

Nowadays, wireless spectrum is becoming increasingly congested and scarce with the

explosive development of wireless devices and services. Dynamic spectrum access (DSA) is a promising approach to increase the spectrum efficiency and alleviate the spectrum scarcity, by allowing unlicensed secondary users (SUs) to opportunistically access the spectrum licensed to primary users (PUs) (Zhao and Sadler 2007; Akyildiz et al. 2008). Hence, it has become one of the important information and communication technologies (ICT) for future wireless networks. To successfully implement DSA in practice, it is important to offer necessary incentives for PUs to open their spectrum for SUs' utilizations and for SUs to access the PUs' spectrum despite the potential costs (Niyato and Hossain 2008; Gao et al. 2011a, b, 2013; Huang et al. 2006; Huang 2013; Luo et al. 2015).

Cooperative spectrum sharing (CSS) has been recently proposed as an effective approach to offer necessary incentives for both PUs and SUs in DSA (Simeone et al. 2008; Zhang and Zhang 2009; Wang et al. 2010; Duan et al. 2011, 2014). The key idea is to enable SUs to act as relay and help the transmission of PUs in exchange for the opportunities of accessing the PUs' licensed spectrum. In such a way, PUs can increase the transmission efficiency with the help of SUs (and hence can save some spectrum resource for SUs), and SUs can obtain the desired transmission opportunity on the PUs' licensed spectrum (at the cost of consuming energy resource to help PUs). Thus, it will lead to a win-win situation for PUs and SUs.

Existing work on CSS mainly focused on the interactions between *one* PU and one or multiple SUs (Simeone et al. 2008; Zhang and Zhang 2009; Wang et al. 2010; Duan et al. 2011, 2014). The PU, as the monopolist, can dominate the cooperation process. Some works assumed that the PU has complete information of SUs and formulated the problem into Stackelberg games (Simeone et al. 2008; Zhang and Zhang 2009; Wang et al. 2010), while others considered that the PU has limited information about SUs and introduced auction theory and contract theory to elicit the private information of SUs (Duan et al. 2011, 2014). However, these models and approaches are difficult to be extended to the sce-

nario of multiple PUs and multiple SUs, which is more general and practical.

This entry focuses on discussing the general CSS model with multiple PUs and SUs. Although the multi-PU multi-SU scenario is more general and practical, it is also much more challenging to analyze the interactions between multiple PUs and multiple SUs, as not only SUs compete with each other for the PUs' licensed spectrums, but PUs also compete with each other for the SUs' collaborations.

Foundations

In a multi-PU multi-SU model, the first key challenging problem is to decide *which PU cooperates with which SU*. Note that multiple SUs may be interested in cooperating with the same PU, and so are PUs. Thus, competition inherently exists not only among PUs but also among SUs. This is a typical *two-sided matching* problem (Gale and Shapley 1962). The second challenging problem is to decide *how each pair of matched PU and SU cooperate with each other*. That is, how much spectrum resource that the PU would like to share with the SU, and how much effort (e.g., energy) that the SU would like to put to help the PU. In this sense, the PU contributes spectrum resource in exchange for the SU's energy resource, or equivalently, the SU contributes energy resource in exchange for the PU's spectrum resource. Thus, this can also be viewed as *resource exchange* between them.

Combining the above two problems, the interaction between multiple PUs and SUs can be formulated as an extended two-sided matching problem with varying resource exchange. In this section, the general CSS model, the extended two-sided matching formulation, and the conditions for stable matching outcome (equilibrium) will be discussed in detail.

General CSS Model

A general CSS model consists of multiple PUs and multiple SUs. Let $\mathcal{M} \triangleq \{1, \dots, M\}$ denote the set of PUs and $\mathcal{N} \triangleq \{1, \dots, N\}$ denote the set of SUs. Each PU owns a dedicated licensed

spectrum band, and the spectrum bands of different PUs are *nonoverlapping* (i.e., orthogonal). Thus, there is no interference on different PUs' spectrum bands. Each SU is equipped with one radio frequency module and hence can access one PU's spectrum band at a particular time.

By cooperative spectrum sharing, SUs act as relays to help the transmissions of PUs in exchange for accessing the PUs' spectrum bands. Assume that each PU can employ *only one* SU as cooperative relay and each SU can help *only one* PU at a particular time. Note that in practice, a PU can possibly employ multiple SUs as cooperative relays. However, existing study Yi et al. (2010) has shown that choosing one relay (i.e., the most appropriate one) is usually sufficient to achieve all or almost all cooperative gain. Similarly, in practice, an SU can possibly help multiple PUs by using a proper multiple access technique. From the economic perspective, however, there is no incentive for an SU to help multiple PUs, as he can only access one PU's frequency band at a particular time.

With the above assumption, the assignment between PUs and SUs is essentially a *one-to-one matching* between them, where one PU (or SU) can only be matched with one SU (or PU). Let μ_n denote the PU matched with SU n and μ_m denote the SU matched with PU m . Note that $\mu_n = \emptyset$ or $\mu_m = \emptyset$ denotes that SU n or PU m is not matched with any PU or SU. Clearly, $\mu_n = m$ if and only if $\mu_m = n$.

For a pair of matched PU and SU, their cooperation frame is divided into 2 phases: in Phase I, the PU transmits his own data with the cooperative relay of the SU; in Phase II, the SU transmits his data on the PU's frequency band (and the PU stops his own data transmission). The cooperative relay in Phase I can use any cooperative protocol such as amplified-and-forward (AF) and decode-and-forward (DF). Obviously, the cooperative gain for the PU greatly depends on the relay effort of the SU, e.g., the SU's transmission power for relaying. Moreover, the cooperative gain for the SU depends on the length of the spectrum access time (Phase II) shared by the PU. As discussed before, such a cooperation process can be viewed as *resource exchange* between them,

where the SU contributes his energy resource (in Phase I) in exchange for using the PU's spectrum resource (in Phase II).

Clearly, a larger relay effort of the SU can bring more benefit for the PU but will also incur higher cost on the SU himself, as he needs to consume more resource for helping the PU. Similarly, a larger spectrum access time shared by the PU in Phase II can bring more benefit for the SU but will reduce the PU's benefit due to the reduced time for his own transmission in Phase I. Thus, for a pair of matched PU and SU, increasing the benefit for one user will accordingly reduce the benefit for the other user. This process can be viewed as the *division of cooperative gain* between them. Let $f_n^m(x)$ denote the *gain division function* (GDF) between PU m and SU n , which characterizes the maximum gain that a PU m can achieve when cooperating with SU n and meanwhile leaving a gain of x for SU n . Let $g_n^m(y)$ denote the *inverse gain division function* (IGDF) between PU m and SU n , which characterizes the maximum gain that an SU n can achieve when cooperating with PU m and meanwhile leaving a gain of y for PU m . Obviously, IGDF is the inverse function of GDF, i.e., $y = f_n^m(x)$ if and only if $x = g_n^m(y)$.

Note that the GDF or IGDF in this entry is quite general and can be arbitrary decreasing function. An explicit example of GDF is provided in Gao et al. (2017), which is defined as the maximum data rate that PU m can achieve by jointly optimizing the SU's relay power in Phase I and spectrum access time in Phase II.

Matching Formulation

Based on the above discussion, the interactions between PUs and SUs can be formulated as an extended *two-sided matching problem*, with PUs on one side and SUs on the other side. A PU is matched with an SU (or equivalently, an SU is matched with a PU) means that the PU cooperates with the SU under certain cooperative gain division (or resource exchange) agreement between them. Accordingly, a *matching outcome* defines not only the cooperative relationships of all PUs and SUs (i.e., who cooperates with whom), but also the cooperative gain division

(or resource exchange) between each pair of matched PU and SU (i.e., how they cooperate with each other). Formally, such a matching outcome can be expressed as

$$\{(\mu_n, \delta_n), \forall n \in \mathcal{N}\}, \quad (1)$$

where (μ_n, δ_n) denotes the matching result for each SU $n \in \mathcal{N}$, with μ_n denoting his matched PU and δ_n denoting his achieved cooperative gain. Obviously, in such a matching outcome, the (maximum) cooperative gain that PU μ_n can achieve is $f_n^{\mu_n}(\delta_n)$. Note that if SU n is not matched with any PU, i.e., $\mu_n = \emptyset$, he cannot achieve any cooperative gain, i.e., $\delta_n = 0$.

Note that a matching outcome may not be stable, as some users may discard their matched partners or change the gain division (resource exchange) agreement with their partners. For example, a matching outcome is not stable, if there exist a PU and an SU who are not matched with each other, but both prefer to be (i.e., both can increase their gains by matching with each other). In such a case, they would match with each other by discarding their respective partners. On the other hand, a matching outcome is also not stable, if there exist a PU and an SU who are matched with each other, but one or both prefer not to be (i.e., one can increase his gain by not matching with the other). In such a case, the user would discard his partner and match with another user or remain single.

Therefore, a natural question arising in such an extended two-sided matching problem is *what are the stable matching outcomes, from which none of PUs and SUs has the incentive to deviate*. A stable matching outcome is often referred to as an *equilibrium* of the matching. Formally,

Definition 1 (Equilibrium) An equilibrium of a matching is a stable matching outcome, where no user can improve his benefit via the unilateral deviation (i.e., choosing another partner or changing the gain division in agreement with his partner).

In the next section, the necessary and sufficient conditions for stable matching outcome (equilibrium) will be derived systematically.

Stable Matching Outcome (Equilibrium)

Suppose $\{(\mu_n, \delta_n), \forall n \in \mathcal{N}\}$ is an equilibrium. Then, for each pair of matched SU and PU (say, SU n and PU μ_n), the following conditions must hold: (i) both SU n and PU μ_n have no incentive to break the current matching, by pairing with another user or remaining single, and (ii) all PUs other than μ_n (or all SUs other than n) have no incentives to discard their respective partners and pair with SU n (or PU μ_n). This leads to the following necessary conditions for equilibrium.

Lemma 1 (Necessary Conditions)

If $\{(\mu_n, \delta_n), \forall n \in \mathcal{N}\}$ is an equilibrium, then for each PU μ_n (matched with SU n), the following condition holds:

$$\begin{aligned} \text{(IR)} \quad & f_n^{\mu_n}(\delta_n) \geq 0, \\ \text{(IC)} \quad & f_n^{\mu_n}(\delta_n) \geq f_k^{\mu_n}(\delta_k), \quad \forall k \neq n, \\ \text{(CC)} \quad & f_{\mu_m}^m(\delta_{\mu_m}) \geq f_n^m(\delta_n), \quad \forall m \neq \mu_n. \end{aligned} \quad (2)$$

(where $f_n^m(x)$ is the GDF, denoting the (maximum) gain that PU m can achieve when matching with SU n and leaving a gain of x for SU n).

The first condition is called the *Individual Rationality* (IR) condition, the second condition is called the *Incentive Compatibility* (IC) condition, and the third condition is called the *Competitive Compatibility* (CC) condition. The IC and IR conditions ensure that each PU μ_n (matched with SU n) has no incentive to discard his current partner SU n by remaining single or matching with another SU. The CC condition ensures that all PUs other than μ_n have no incentive to discard their current partners and compete with PU μ_n for SU n .

The IR, IC, and CC conditions lead to the following constraints on δ_n :

$$\begin{aligned} \text{(IR)} \quad & \Rightarrow \delta_n \leq g_n^{\mu_n}(0), \\ \text{(IC)} \quad & \Rightarrow \delta_n \leq \min_{k \neq n} \{g_n^{\mu_n}(f_k^{\mu_n}(\delta_k))\}, \\ \text{(CC)} \quad & \Rightarrow \delta_n \geq \max_{m \neq \mu_n} \{g_n^m(f_{\mu_m}^m(\delta_{\mu_m}))\}, \end{aligned} \quad (3)$$

where $g_n^m(x) = f_n^{\mu_n(-1)}(x)$ denotes maximum gain that SU n can achieve when matching with PU m and leaving a gain of x for PU m . The above

constraints on δ_n further lead to the following upper-bound and lower-bound for each δ_n :

$$\bar{\delta}_n \triangleq \min_{k \neq n} \{g_n^{\mu_n}(f_k^{\mu_n}(\delta_k)), g_n^{\mu_n}(0)\}. \quad (4)$$

$$\underline{\delta}_n \triangleq \max_{m \neq \mu_n} \{g_n^m(f_{\mu_n}^m(\delta_{\mu_n})), 0\}, \quad (5)$$

Intuitively, the gain for SU n , i.e., δ_n , cannot be higher than the threshold $\bar{\delta}_n$, otherwise PU μ_n (matched with SU n) would be better off by matching with another SU or remaining single. On the other hand, δ_n cannot be lower than the threshold $\underline{\delta}_n$, otherwise some other PUs $m \neq \mu_n$ (matched with other SUs) will have the incentive and ability to compete for SU n (by leaving a slightly higher gain for SU n).

With the above upper-bound and lower-bound for each δ_n , the necessary conditions in Lemma 1 can be rewritten as $\underline{\delta}_n \leq \delta_n \leq \bar{\delta}_n, \forall n \in \mathcal{N}$. It is important to note that these conditions are not only necessary, but also sufficient. Formally,

Lemma 2 (Necessary and Sufficient Conditions)

A matching outcome $\{(\mu_n, \delta_n), \forall n \in \mathcal{N}\}$ is an equilibrium, if and only if:

$$\underline{\delta}_n \leq \delta_n \leq \bar{\delta}_n, \quad \forall n \in \mathcal{N}, \quad (6)$$

where $\bar{\delta}_n$ and $\underline{\delta}_n$ are defined in Eqs. (4) and (5), respectively.

Key Applications

Two-sided matching is a widely used efficient framework for analyzing the interactions between two disjoint player sets in two-sided market scenarios. In such scenarios, the matching theory can systematically capture not only the cooperative interactions between users in different sides but also the competitive interactions between users on the same side. Most early results in this area mainly focused on the matching under complete information, without considering the incentive issues under incomplete information.

The first two-sided matching model was studied in Gale and Shapley (1962), where a basic two-sided matching model with nontransferable utilities was proposed to study the marriage market. Shapley and Shubik (1971) and Thompson (1980) studied the more general models with additive and transferable utilities. Kelso and Crawford (1982) studied the two-sided labor models with nontransferable utilities. Some later works such as Kojima and Pathak (2009) studied the incentive issue in matching under incomplete information. Gao et al. (2017) applied the matching theory to analyze the dynamic spectrum access problem between PUs and SUs in cognitive radio networks.

Open Problems

In this section, some open problems for the applications of matching to communication networks are outlined.

- *One-to-Many or Many-to-Many Matching:* This entry mainly focused on the one-to-one matching model, where a PU can only employ one SU as cooperative relay and an SU can only help one PU. In practice, however, a PU can potentially employ multiple SUs as relays and an SU can potentially help multiple PUs. This will lead to a more complicated one-to-many or many-to-many matching.
- *Implementation of Equilibrium:* This entry mainly focused on characterizing the conditions of equilibrium, without considering the implementation of equilibrium under different information scenarios, especially the incomplete information scenario. That is, what equilibrium will emerge under different information scenarios is a challenging open problem.

Cross-References

- [Efficiency and Pareto Optimality](#)
- [Matching Game for Load Distribution in Multi-tier Cellular Networks](#)
- [Preferences and Utility Functions](#)

References

- Akyildiz I, Lee W, Vuran M, Mohanty S (2008) A survey on spectrum management in cognitive radio networks. *IEEE Commun Mag* 46:40–48
- Duan L, Gao L, Huang J (2011) Contract-based cooperative spectrum sharing. In: *IEEE symposium on new frontiers in dynamic spectrum access networks (DySPAN)*, pp 399–407
- Duan L, Gao L, Huang J (2014) Cooperative spectrum sharing: a contract-based approach. *IEEE Trans Mob Comput* 13:174–187
- Gale D, Shapley L (1962) College admissions and the stability of marriage. *Am Math Mon* 69:9–15
- Gao L, Wang X, Xu Y, Zhang Q (2011a) Spectrum trading in cognitive radio networks: a contract-theoretic modeling approach. *IEEE J Sel Areas Commun* 29: 843–855
- Gao L, Xu Y, Wang X (2011b) MAP: multi-auctioneer progressive auction for dynamic spectrum access. *IEEE Trans Mob Comput* 10:1144–1161
- Gao L, Huang J, Chen Y, Shou B (2013) An integrated contract and auction design for secondary spectrum trading. *IEEE J Sel Areas Commun* 31: 581–592
- Gao L, Duan L, Huang J (2017) Two-sided matching based cooperative spectrum sharing. *IEEE Trans Mobile Comput* 16:538–551
- Huang J (2013) Market mechanisms for cooperative spectrum trading with incomplete network information. *IEEE Commun Mag* 51:201–207
- Huang J, Berry R, Honig M (2006) Auction-based spectrum sharing. *Mobile Netw Appl* 11:405–418
- Kelso JA, Crawford V (1982) Job matching, coalition formation, and gross substitutes. *Econometrica* 50:1483–1504
- Kojima F, Pathak P (2009) Incentives and stability in large two-sided matching markets. *Am Econ Rev* 34:383–387
- Luo Y, Gao L, Huang J (2015) Business modeling for TV white space networks. *IEEE Commun Mag* 53:82–88
- Niyato D, Hossain E (2008) Spectrum trading in cognitive radio networks: a market-equilibrium-based approach. *IEEE Wirel Commun* 15:71–80
- Shapley L, Shubik M (1971) The assignment game I: the core. *Int J Game Theory* 1:111–130
- Simeone O, Stanojev I, Savazzi S, Bar-Ness Y, Spagnolini U, Pickholtz R (2008) Spectrum leasing to cooperating secondary ad hoc networks. *IEEE J Sel Areas Commun* 26:203–213
- Thompson GL (1980) Computing the core of a market game. In: *Fiacco A V, Kortanek K O (eds) Extremal Methods and Systems Analysis. Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, 174:312–334
- Wang H, Gao L, Gan X, Wang X, Hossain E (2010) Cooperative spectrum sharing in cognitive radio networks: a game-theoretic approach. In: *IEEE international conference on communications (ICC)*, pp 1–5
- Yi Y, Zhang J, Zhang Q, Jiang T, Zhang J (2010) Cooperative communication-aware spectrum leasing in cognitive radio networks. In: *IEEE symposium on new frontiers in dynamic spectrum access networks (DySPAN)*, pp 1–11
- Zhang J, Zhang Q (2009) Stackelberg game for utility-based cooperative cognitive radio networks. In: *ACM international symposium on mobile ad hoc networking and computing (MobiHoc)*, pp 3412–3415
- Zhao Q, Sadler M (2007) A survey of dynamic spectrum access. *IEEE Signal Process Mag* 24:79–89

Matching Game for Load Balancing in Wireless Cellular Networks

► Matching Game for Load Distribution in Multi-tier Cellular Networks

Matching Game for Load Distribution in Multi-tier Cellular Networks

Nima Namvar¹ and Behrouz Maham²

¹Department of Electrical and Computer Engineering, North Carolina A&T State University, Greensboro, NC, USA

²Electrical and Electronic Engineering Department, Nazarbayev University, Astana, Kazakhstan

Synonyms

Matching game for load balancing in wireless cellular networks

Definitions

Load distribution in wireless cellular networks refers to assigning the mobile users (MUs) to the base stations (BSs) throughout the network. Multi-tier cellular networks (MTCNs) are a kind of wireless cellular networks where the access points come in various size, capacity,

and transmit power. Load distribution in such networks should take into account the disparity among the access points so as to optimize the network performance.

Introduction

The concept of multi-tier cellular network (MTCN) is seen as a promising technology to cope with the unprecedented proliferation in wireless data traffic by increasing the network capacity (Damjanovic et al. 2011). In MTCN, traditional macrocells coexist with other smaller cells, such as microcells and femtocells, which are the catalyst behind improving wireless networks in densely populated areas. In comparison to their macrocell counterparts, small cell transmissions range between a mere 0.25 and 6 W and extend coverage up to several hundred meters. Hence, by reducing the distance between the transmitter and the receiver, small cells provide higher data rates as well as more energy-efficient way of communication for their users. However, the deployment of MTCN comes with several new technical challenges, such as interference management, network planning, and load distribution (Ghosh et al. 2012). In particular, load distribution strategy in MTCN is of crucial importance for the overall network performance (Ye et al. 2013), as it determines which user should be served by which BS and when. In this article, we propose a novel strategy for load distribution in the downlink of MTCNs by exploiting a combination of new context information related to the users trajectory profile, their QoS demands, and the current load of each cell.

Problem Formulation

Consider the downlink of a two-tier MTCN consisting of macrocells and picocells. Let $\mathcal{M}, \mathcal{P}$, and $\mathcal{N}$ denote the set of M macro base stations (MBSs), the set of P pico base stations (PBSs), and the set of N users, respectively. For each user $n \in \mathcal{N}$, we associate a unique profile $P_n(\tau, \theta, V)$

defined by three variables τ, θ , and V representing the urgency of the service in use, the direction of its motion, and its velocity, respectively. Next, we study these parameters, and we show how such information can improve the load distribution strategy.

Service Urgency

Depending on their application in use, the users may have varying needs for the wireless resources and tolerate different amounts of latency. To capture the dependency of the users' quality of service (QoS) to the service latency, we define the users' QoS as a decreasing function of the delivery time similar to Proebster et al. (2011):

$$Q_n(t) = \frac{1}{1 + \exp(t - \tau_n)}, \quad (1)$$

where τ_n is the urgency coefficient, which accounts for the urgency of the data. The smaller is τ_n , the more urgent is the data in use.

Users' Trajectory

Users travel through the network, and active communication sessions must be transferred among the cells. To guarantee the QoS, the network should avoid risky handovers that may lead to a signal loss or erroneous communication. A handover fails when the received SINR drops under a certain threshold. To capture the effect of the users' trajectory on the user-cell association problem, we derive the probability of handover failure when a user moves between different cells.

Without loss of generality, we assume circular coverage area for tractability (Lopez-Perez et al. 2012). Consider two picocells in the vicinity of each other and assume that a user is traveling through these two cells. When a user enters a cell, the total possible time of interaction is $T_i = \frac{D}{V}$, where D is the length of coverage circle chord traversed by the user and V is its speed. Also, assume that a successful handover process needs a certain preparation time of duration T_p . If $T_i > T_p$, the user is considered as a candidate to be served; otherwise, no handover would be initiated. Note

that $D = 2R \cos(\theta)$ where R is the radius of the coverage circle and θ is the trajectory angle measured with respect to the horizontal line in the polar coordinate system. Since the directions which the user takes into the picocell are equally likely, θ is modeled as a uniform random variable $\theta \sim \text{Uniform}(-\frac{\pi}{2}, \frac{\pi}{2})$. Hence, the cumulative distribution function (CDF) of D is given by:

$$\begin{aligned} \Pr(D < d) &= 2\Pr\left(\theta > \cos^{-1}\left(\frac{d}{2R}\right)\right) \\ &= 1 - \frac{2}{\pi} \cos^{-1}\left(\frac{d}{2R}\right). \end{aligned} \quad (2)$$

We assume that for having a successful handover, the handover process must be done before the user reaches the distance $r < R$ of the picocell. When the user's path tangent to the circle with radius r around the picocell with coverage radius R , D is equal to $2\sqrt{R^2 - r^2}$. Handover fails when the user's path intersects the circle r ; thus the probability of handover failure is:

$$\begin{aligned} \Pr(D \geq 2\sqrt{R^2 - r^2}) \\ = \frac{2}{\pi} \cos^{-1}\left(\sqrt{1 - \left(\frac{r}{R}\right)^2}\right). \end{aligned} \quad (3)$$

The expression in (3) shows that the handover process becomes more reliable as r becomes smaller relative to R . The ratio of r to R varies for each cell, and thus, the different cells guarantee different levels of reliability in handover process.

User-Cell Association as a Matching Game

Matching theory is a mathematical framework for modeling the combinatorial problem of matching players in two distinct sets, depending on the individual information and preference of each player (Gu et al. 2015). Here, we model the problem of user-cell association in MTCNs as a many-to-one matching game between the cells and the users.

Definition 1 A matching μ is a function from $\mathcal{N} \cup \mathcal{P}$ to $2^{\mathcal{N} \cup \mathcal{P}}$ such that $\forall n \in \mathcal{N}$ and $\forall p \in \mathcal{P}$: (i) $\mu(n) \in \mathcal{P}$ and $|\mu(n)| \leq 1$, (ii) $\mu(p) \in 2^{\mathcal{N}}$ and $|\mu(p)| \leq q_p$ where q_p is the quota of p , and (iii) $\mu(n) = p$ if and only if n is in $\mu(p)$.

The users, who are not assigned to any member of $\mathcal{P}$, will be assigned to the nearest macro BS. Members of $\mathcal{N}$ and $\mathcal{P}$ must have strict, reflexive, and transitive preferences over the agents in the opposite set. Next, exploiting the context information about the users' trajectory and service urgency, we introduce some properly defined utility functions to effectively capture the preferences of each set.

Users' Preferences

Users demand for reliable and high-quality communication. Therefore, they prefer the SCBSs which can guarantee the safety of communication during the handover and are able to deliver the data in an acceptable time. The transmission rate a user receives from a picocell depends on the current cell load and the interference caused by the neighboring cells. Therefore, the rate of transmission is a function of current matching μ . We define the rate over load for all user-cell pairs as:

$$\begin{aligned} \nabla_{ij} &= \frac{1}{\max(1, K_j)} \\ &\log_2 \left(1 + \frac{P_j c_{ij}}{\sum_{k \neq j} P_k c_{ik} + \sigma^2} \right), \end{aligned} \quad (4)$$

where K_j is the total users being served by picocell j and P_j denotes its power. c_{ij} represents the channel coefficient between user i and picocell j . In addition, σ^2 is the power of additive noise. We define the following utility function for user i when admitted by picocell j :

$$U_i(\mu, R_j, r_j, \nabla_{ij}) = \frac{R_j}{r_j} \nabla_{ij}. \quad (5)$$

The expression in (5) shows that the users prefer lightly loaded picocells to maximize their utility. Moreover, note that this utility function could

help to offload the heavily loaded macrocells by pushing the users to more lightly loaded cells.

Picocells' Preferences

The main goal of picocells is to increase the network-wide capacity by offloading traffic from the macrocells while providing satisfactory QoS for the users. Each picocell p could serve a limited number of users, called its quota q_p . Thus, by prioritizing the users coming from congested cells, the picocells could offload the heavily loaded cells. On the one hand, each candidate user is carrying a potential utility as a function of the pervious cell p' load, i.e., $f\left(\frac{K_{p'}}{q_{p'}}\right)$. This utility depends on the current matching which determines the number of users in neighboring cells. On the other hand, picocell p would also prioritize the candidate users based on their service requirement, i.e., service urgency defined in (1). We define the following utility that picocell p obtains by serving a user n in its coverage area:

$$U_p(\mu, \tau_n, k_{p'}, q_{p'}) = \left[1 + \log \left(\frac{\max(1, k_{p'})}{q_{p'}} \right) \right] \frac{1}{\tau_n}. \quad (6)$$

The first term in (6) accounts for the offloading concept, and the second term is the utility achieved by the picocells p for its service to a specific application. Notice that the small cell p gains more utility by giving service to the users having more urgent data. Using these utility functions, users build an ordered list of the picocells based on their own preferences and vice versa. Here we define the preference relationship for the users as follows:

Definition 2 The preference relation $\succ_n$ of the user $n \in \mathcal{N}$ over the set $\mathcal{P}$ is a function that compare two picocells p and p' such that:

$$\mu \succ_n \mu' \Leftrightarrow U_n(\mu) > U_n(\mu'). \quad (7)$$

The preference relation for the picocells is defined similarly. Users and picocells rank the members of the opposite set based on the defined preference relations. Our purpose is to match the users to the small cells so that the preferences of both sides will be satisfied as much as possible; thereby the network-wide efficiency would be optimized. To solve a matching game, one suitable concept is that of a stable matching (Gu et al. 2015). A matching is said to be two-sided stable, if and only if there is no blocking pair, i.e., there is no user-picocell pair (n, p) where n prefers p to its currently matched user picocell and p prefers n to its currently matched user n .

Next, an efficient algorithm for solving the matching problem is presented, which reaches to a stable matching between users and picocells.

Proposed Algorithm

The deferred acceptance algorithm is a well-known approach to solving the standard matching games (Roth and Sotomayo 1992). However, in our game, the preferences of agents as shown in (5) and (6) depend on externalities through the entire matching, unlike classical matching problems in which preferences are static and independent of the matching. Therefore, the classical approaches such as the deferred acceptance cannot be used here because of the presence of externalities. To solve the formulated game, we propose a novel algorithm shown in Table 1.

Suppose that all the users are initially associated to the nearest macro BS. Each user sends its profile information to the neighboring picocells. Each picocell, on the other side, ranks the candidate users based on its utility and feeds back the awaiting users with its own context information including its rate over load defined in (4) and its corresponding coverage and handover circle radii R and r . Each user makes a ranking list of the available picocells and applies to the most preferred one of them. The picocells keep the most preferred ones up to their quota and reject the others. The users who have been rejected in the former phase would apply to their next

Matching Game for Load Distribution in Multi-tier Cellular Networks, Table 1

Proposed algorithm for the user-cell association game

Input: context-aware utilities and the preferences of each user

Output: Stable matching between the D2D pairs

Initializing: All the UEs are assigned to the nearest BS

Stage I: Preference Lists Composition

- Neighboring D2D users exchange their context information
- Users sort the set of acceptable candidate based on their preference functions

Stage II: Matching Evaluation

while: $\mu^{(n+1)} \neq \mu^{(n)}$

- Update the utilities based on the current matching μ
- Construct the preference lists using preference relations
- Each user n applies to its most preferred partner
- Each user accept the most preferred applicants and create a waiting list while rejecting the others

Repeat

- Each rejected user applies to its next preferred partner
- Each user update its waiting list considering the new applicants

Until: all the users assigned to a waiting list

end

M

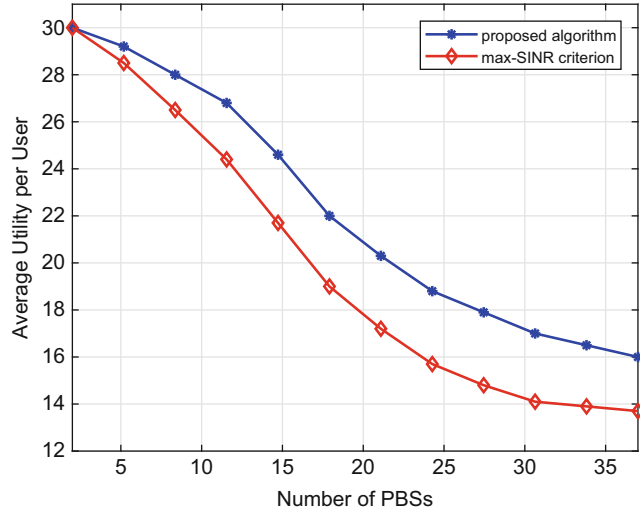
favorite picocell, and the picocells modify their waiting list accordingly. This procedure continues until all the users are assigned to a waiting list. However, since the preferences depend on the current matching μ , an iterative approach should be employed. In each step, the utilities would be updated based on the current matching. Once the utilities are updated, the preference lists would be updated accordingly as well. Therefore, in each iteration, a new temporal matching arises, and based on this matching, the interdependent utilities are updated as well. The algorithm initiates the next iteration based on the modified preferences. The iterations run on until two subsequent temporal matchings are the same and algorithm converges. We note that the proposed algorithm will lead to a stable matching when it converges. Indeed, the deferred acceptance in stage *II* would not converge if the matching is not stable. Hence, by contradiction, whenever the algorithm converges, the matching would be stable.

Simulation Results

For our simulations, we consider a single MBS with radius 1 km and overlaid by P uniformly deployed picocells. The channels suffer from a Rayleigh fading with parameter $\sigma = 2$. Noise level is assumed to be $\sigma^2 = -121$ dBm, and the minimum acceptable SINR for the UEs is 9.56 dB. There are N users distributed uniformly in the network. The QoS parameter τ_n is chosen randomly from the interval $[0.5, 5]$ ms. The users have low mobility and can be assumed approximately static during the process time required for a matching. All the statistical results are averaged by 1000+ runs over random location of users and SCBSs, the channel fading coefficients, and other random parameters. The performance is compared with the max-SINR algorithm which is a well-known context-unaware approach exploited in wireless cellular networks for user-cell association. In this approach, each user is associated to the SCBS providing the strongest SINR.

Matching Game for Load Distribution in Multi-tier Cellular Networks, Fig. 1

Average utility per user for different number of PBSs with $N = 60$ users



Matching Game for Load Distribution in Multi-tier Cellular Networks, Fig. 2

Average utility per PBS for different number of users with $P = 20$ PBSs

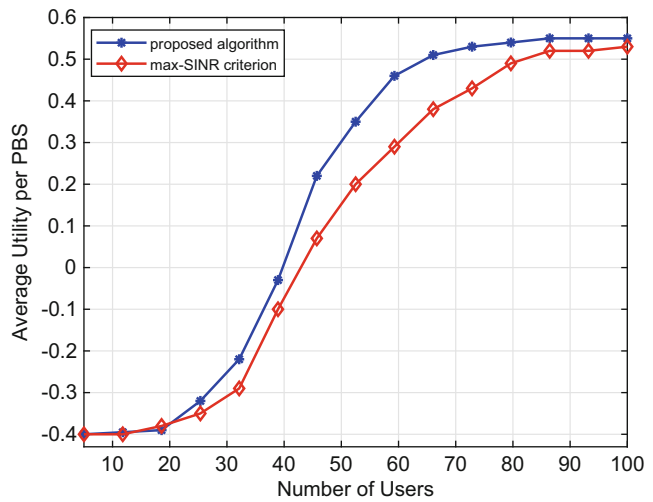


Figure 1 shows the average utility per user as a function of the number of PBSs for $N = 60$ users. As the number of PBSs increases, the average load in each cell decreases. Consequently, the available resources for each user would increase. However, increasing the number of PBSs leads to an increase in the interference between the neighboring cells and, consequently, decreases the rate. Hence, as shown Fig. 1, the average utility per user is a decreasing function with respect to the number of PBSs. We can also see that, when the number of PBSs exceeds 30, the slope of the curves becomes less steep due to the fact that the matchings would not be affected by these increments as much as before. Indeed,

beyond this point, there are enough PBSs to serve all the users, and deploying more PBSs will not change the current matching dramatically. Figure 1 demonstrates that the proposed context-aware approach has a considerable performance advantage compared to the max-SINR approach. This performance advantage reaches up to 20.4% gain over to max-SINR criterion for a network with 30 PBSs.

Figure 2 shows the average utility achieved by each PBS as a function of the number of users for $P = 20$ PBSs. As the number of users N increases, the network becomes more congested, and the probability that a new user who applies for a PBS is coming from a

congested BS increases. Therefore, it is more likely for the PBSs to gain more utility by offloading the network. However, when the network is considerably congested, the new users that arrive to the network would be mostly assigned to the MBS, since many of PBSs are already servicing their maximum capacity. In this respect, Fig. 2 shows that, once the number of users exceeds 80, the average utility of PBSs remains constant. The proposed algorithm achieves up to 24.9% gain over the max-SINR approach when the number of users is 50.

Conclusions

In this work, we have proposed a new context-aware user association algorithm for the downlink of the multi-tier cellular networks. By introducing well-designed utility functions, our approach accounts for the trajectory and speed of the users as well as for their heterogeneous QoS requirements. We have modeled the problem as a many-to-one matching game with externalities, where the preferences of the players are interdependent and contingent on the current matching. To solve the game, we have proposed a novel algorithm that converges to a stable matching in a reasonable number of iterations. Simulation results have shown that the proposed approach yields considerable gains compared to max-SINR approach.

Cross-References

- [Heterogeneous Wireless Networks Based on Cognitive Radio](#)
- [Interference-Aware Distributed Resource Allocation](#)

References

Damnjanovic A, Montojo J, Wei Y, Ji T, Luo T, Vajapeyam M, Yoo T, Song O, Malladi D (2011) A survey on 3GPP heterogeneous networks. *IEEE Wirel Commun* 18:1021

- Ghosh A, Mangalvedhe N, Ratasuk R, Mondal B, Cudak M, Visotsky E, Thomas TA, Andrews JG, Jo PXHS, Dhillon HS, Novlan TD (2012) Heterogeneous cellular networks: from theory to practice. *IEEE Commun Mag* 50(6):54–64
- Gu Y, Saad W, Bennis M, Debbah M, Han Z (2015) Matching theory for future wireless networks: fundamentals and applications. *IEEE Commun Mag* 53(5):52–59
- Lopez-Perez D, Guvenc I, Xiaoli C (2012) Theoretical analysis of handover failure and ping-pong rates for heterogeneous networks. In: *IEEE international conference on communications (ICC)*, pp 6774–6774
- Proebster M, Kaschub M, Valentin S (2011) Context-aware resource allocation to improve the quality of service of heterogeneous traffic. In: *IEEE international conference on communications (ICC)*, pp 1–6
- Roth AE, Sotomayo MAO (1992) Two-sided matching: a study in game-theoretic modeling and analysis. Cambridge University Press, Cambridge
- Ye Q, Rong B, Chen Y, Al-Shalash M, Caramanis C, Andrews JG (2013) User association for load balancing in heterogeneous cellular networks. *IEEE Trans Wirel Commun* 12(6):2706–2716

Mathematical Model

- [Reaction-Diffusion Channels](#)

Maximum Likelihood Sequence Detection

- [Equalization Techniques for Single-Carrier Modulations](#)

Media Access Control for Narrowband Internet of Things: A Survey

Yuyi Sun, Shibo He, and Fei Tong
The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

Introduction

Low-power wide-area network (LPWAN) has emerged as an attractive communication platform

for Internet of Things (IoT) (Eletreby et al. 2017). Among a variety of communication technologies in LPWAN, narrowband IoT (NB-IoT), standardized in the third generation partnership project in Release 13 (3GPP 2015), is a promising technology that attracts attention from both industry and academia. NB-IoT inherits long term evolution (LTE) and has its distinctive advantages, including (1) massive connections, (2) wide-area coverage, (3) low power consumption, and (4) low cost (Wang et al. 2017; Beyene et al. 2017; Zayas and Merino 2017).

Media access control (MAC) protocol in NB-IoT resolves channel collisions among multiple user equipments (UEs) (3GPP TS 36.211; 3GPP TS 36.331; 3GPP TS 36.321), by efficiently allocating channel resources of the evolved Node Bs (eNBs) to UEs. When there is a data transmission demand, a UE initiates random access procedure to occupy a channel and transmits data to an eNB. Clearly, random access protocol suffers from severe collisions when there are a large number of UEs in NB-IoT, leading to more power consumption. Therefore, it is significant to efficiently analyze the performance of random access in NB-IoT.

In LTE, MAC protocols can be divided into contention-free and contention-based MAC protocols, and here we focus on contention-based MAC protocols, i.e., contention-based random access procedure. Based on random access procedure, there are some existing models designed for LTE, which cannot be applied to NB-IoT directly since there are a large number of UEs in NB-IoT and meanwhile, wide-area is required. Having this in mind, we provide a survey of the existing MAC models for LTE and NB-IoT in this paper.

MAC for LTE

Random Access Procedure in LTE

Random access procedure in LTE consists of four steps: random access preamble, random access response (RAR), scheduled transmission, and contention resolution.

- *Random Access Preamble*: a UE selects one of the preamble sequences from two groups and transmits it to an eNB.
- *RAR*: after the eNB receives the preamble sequence, it transmits an RAR to the UE. If more than one UE selects the same preamble sequence, they will receive the same RAR.
- *Scheduled Transmission*: after receiving the RAR, the UE transmits/UEs transmit a message that contains a unique identity.
- *Contention Resolution*: if the eNB can decode one of the messages from scheduled transmission, it replies to the UE to transmit data.

MAC Models for LTE

Data transmission, power mechanism, and energy consumption are crucial to MAC for LTE. Previous studies have proposed different models to analyze such problems.

Markov chain was adopted to model random access procedure and analyze data transmission in LTE (Seo and Leung 2012). Based on semi-persistent scheduling (SPS) in random access for LTE, UEs are assumed to have an infinite buffer to store packets. Markov chain is used to model the SPS in LTE, including the states ON and OFF of the traffic source, the contention and transmission states of UEs. Equilibrium point analysis is investigated, obtaining the number of UEs in steady state, through which the probability of packet dropping is analyzed.

Power ramping mechanism was modeled during random access in LTE in terms of preamble signal to interference plus noise ratio (SINR) and preamble collisions (Misic et al. 2017). The ratio of external interference to received signal power is assumed as a Gaussian distribution, and the probability of missed preamble detection and the failure caused by eNB outage can be analyzed. Performance evaluation reveals that the downlink resources bring more success probability.

Zhao et al. (2017) reduced power consumption in a random access algorithm based on statistic waiting. After investigating random access procedure in LTE, they used slotted ALOHA to simulate packet transmission. A UE initializing random access procedure receives the success

rate of the last time slot, which helps the UE decide whether to transmit a packet or not and reduces energy consumption.

Previous MAC models in LTE cannot be directly applied to NB-IoT due to the following reasons: (a) there are three coverage levels defined by minimum coupling loss (MCL) in NB-IoT (Ng et al. 2009), which covers a wider range than that in LTE, requiring more retransmission times to achieve wide-area coverage. (b) Due to massive connected UEs in NB-IoT, a more efficient MAC model is needed to resolve a great deal of collisions. (c) Due to low-traffic rate and low-power consumption requirements, UEs in NB-IoT have a finite buffer to cache data packets, while LTE requires higher data rate. Thus, the length of buffer and larger retransmission number need to be considered simultaneously in NB-IoT to assign channel resources. We introduce the MAC models for NB-IoT in the next section.

MAC for NB-IoT

Random Access Procedure in NB-IoT

Random access procedure in NB-IoT also consists of four steps: random access preamble, RAR, scheduled transmission, and contention resolution.

- *Random Access Preamble*: a UE transmits a preamble to an eNB, which has a cyclic prefix and five symbols 1. Physical resources are configured based on MCL.
- *RAR*: after the eNB receives the preamble sequence, it transmits an RAR to the UE. If more than one UE transmits the same preamble sequence, they will receive the same RAR.
- *Scheduled Transmission*: after receiving the RAR, the UE transmits/UEs transmit a message that contains a unique identity, power headroom report, buffer states, and data volume.
- *Contention Resolution*: the eNB randomly selects a UE to decode its information received from scheduled transmission, then it replies to the UE to transmit data.

MAC Models for NB-IoT

Previous MAC models for NB-IoT mainly focus on coverage requirement and system throughput.

Lin et al. (2016) designed a frequency hopping pattern for NB-IoT with single tone random access channel signal. Based on the rationale of random access channel frequency hopping, the authors proposed corresponding receiver algorithms and time-of-arrival (ToA) estimation. Comparing the statistic value of ToA and a predefined threshold can help the eNB determine the presence of the preamble sequence. Simulation results reveal that the detection probabilities exceed 99%, which can meet the coverage requirement.

Based on probabilities characterization, first-in-first-out (FIFO) queue model, and Markov chain, a system throughput model was proposed and analyzed (Sun et al. 2017). Based on random access procedure, the UEs' buffer is denoted as a FIFO queue, and the probability that a FIFO queue is empty, the probability that a packet is transmitted successfully, and the probability that a channel is busy are characterized based on backoff mechanism. Markov chain is adopted to model retransmission number and queue length simultaneously to achieve the three probabilities in steady state. The expression of the system throughput can be obtained and calculated due to the three probabilities above. Extensive simulation results show that the number of UEs and the arrival rate of packets have obvious effects on system throughput, while the maximum retransmission number and the packet length have slight influences on the system throughput.

Harwahu et al. (2018) investigated a promising optimized random access model for NB-IoT. NB-IoT systems have three coverage enhancement (CE) levels due to different MCL. A UE initializes random access to occupy a channel from its initial CE level, which depends on the reference signal received power. When collisions happen, if the UE does not transmit successfully till the maximum retransmission number in this CE level is reached, it will switch to the next higher CE level to retransmit. When the global maximum retransmission number is reached, the UE will initialize random access again from its

initial CE level. Thus, in this entry, multiband multichannel slotted ALOHA is used to model this procedure. The first transmission and the retransmission at the same CE level and in the next higher CE level are analyzed. The optimization of random access for NB-IoT in this entry includes that: (a) more CE levels bring larger access success probability, (b) when the number of remaining retransmissions in a certain band is equal to the number of global retransmissions, access success probability increases, and (c) comparing number of successful UEs in each CE level and in their first transmission by exhaustive search can achieve the largest access success probability. Analytical model is supported by extensive simulations. The influences of parameters such as the number of UEs, the repartition ratios of subchannels, and retransmission times for different levels on the access success probability are analyzed.

In Li et al. (2018), an optimization to maximize the system throughput in random access narrowband cognitive radio IoT was introduced. This entry used slotted ALOHA to model random access procedure, including autonomous and collaborative sensing. With unique optimal detection probabilities in network level characterized, the throughput optimization of random access with autonomous sensing and collaborative sensing can be maximized. From the simulations, the system throughput with high signal noise ratio (SNR) and system interference tolerance is higher than that with low SNR and system interference tolerance. Autonomous sensing is better under high SNR, while collaborative sensing has a better performance under low SNR. The arrival rate of data is a Poisson process. Smaller arrival rate of data increases the maximum throughput rapidly and then, the maximum throughput reaches a peak, while larger arrival rate of data makes the maximum throughput decrease. As the arrival rate of data tends to infinity, the probability of detection converges to one, and the optimal probability of false alarm becomes a global optimal problem.

The NB-IoT development system in Chen et al. (2017) includes IoT cloud platform, NB-IoT development board, mobile phones, and appli-

Media Access Control for Narrowband Internet of Things: A Survey, Table 1 Summary of MAC models for NB-IoT

Literature	Design	Analysis
Lin et al. (2016)	Frequency hopping pattern	ToA estimation
Sun et al. (2017)	Finite FIFO queue for UE buffer	System throughput
Harwahu et al. (2018)	Multiband multichannel slotted ALOHA	Access success probability
Li et al. (2018)	Slotted ALOHA for MAC	Throughput optimization
Chen et al. (2017)	Development system	Practical application

cation server. The development board embeds the entire protocol of NB-IoT, including MAC and physical protocols. Based on the applicable system, practical experiments about MAC for NB-IoT can be done, which helps design efficient MAC models for NB-IoT.

Here we summarize the MAC models for NB-IoT in Table 1.

Conclusion

In this entry, we make a survey about the MAC for NB-IoT. Based on random access procedure, there are a lot of MAC models existing for LTE. However, such models cannot be applied to NB-IoT systems due to massive connections and wide-area coverage in NB-IoT. Thus, based on the features of NB-IoT, some works design new models for MAC in NB-IoT. In the future, some more parameters, such as the system latency, packet loss can be analyzed for the system performance of MAC for NB-IoT.

Acknowledgments This work was supported by NSFC under grant 61672458.

References

- 3GPP (2015) New Work Item: NarrowBand IoT (NB-IoT). The 69th SG RAN meeting. Available: https://www.3gpp.org/FTP/tsg_ran/TSGRAN/TSGR69/Docs/RP-151621.zip

- 3GPP TS 36.211, V13.2.0 2016 Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation; Protocol specification (Release 13)
- 3GPP TS 36.321, V13.2.0 2016 Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC); Protocol specification (Release 13)
- 3GPP TS 36.331, V13.2.0 2016 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 13)
- Beyene YD, Jantti R, Tirkkonen O, Ruttik K, Iraj S, Larmo A, Tirronen T, Torsner J (2017) NB-IoT technology overview and experience from cloud-RAN implementation. *IEEE Wirel Commun* 24(3):26–32
- Chen J, Hu K, Wang Q, Sun Y, Shi Z, He S (2017) Narrowband internet of things: implementations and applications. *IEEE Internet Things J* 4(6):2309–2314
- Eletreby R, Zhang D, Kumar S, Yağan O (2017) Empowering low-power wide area networks in urban settings. In: *Proceedings of the conference of the ACM special interest group on data communication*. ACM, Los Angeles, pp 309–321
- Harwahu R, Cheng RG, Wei CH, Sari RF (2018) Optimization of random Access Channel in NB-IoT. *IEEE Internet Things J* 5(1):391–402
- Li T, Yuan J, Torlak M (2018) Network throughput optimization for random access narrowband cognitive radio internet of things (NB-CR-IoT). *IEEE Internet Things J* 5:1436. <https://doi.org/10.1109/JIOT.2017.2789217>
- Lin X, Adhikary A, Wang YPE (2016) Random access preamble design and detection for 3GPP narrowband IoT systems. *IEEE Wireless Commun Lett* 5(6):640–643
- Misic J, Misic, VB, Ali MZ (2017) Explicit power ramping during random access in LTE/LTE-A. In: *Proceedings of IEEE wireless communications and networking conference*, IEEE, San Francisco, pp 1–6
- Ng MH, Lin SD, Li J, Tatesh S (2009) Coexistence studies for 3GPP LTE with other mobile systems. *IEEE Commun Mag* 47(4):60–65
- Seo JB, Leung VCM (2012) Performance modeling and stability of semi-persistent scheduling with initial random access in LTE. *IEEE Trans Wirel Commun* 11(12):4446–4456
- Sun Y, Tong F, Zhang Z, He S (2017) Throughput modeling and analysis of random access in narrowband internet of things. *IEEE Internet Things J* 5:1485. <https://doi.org/10.1109/JIOT.2017.2782318>
- Wang YPE, Lin X, Adhikary A, Grovlen A, Sui Y, Blankenship Y, Bergman J, Razaghi HS (2017) A primer on 3GPP narrowband internet of things. *IEEE Commun Mag* 55(3):117–123
- Zayas AD, Merino P (2017) The 3GPP NB-IoT system architecture for the internet of things. In: *Proceedings of IEEE communications workshops*, IEEE, Paris, pp 277–282
- Zhao Y, Yang H, Liu K, Huang L, Shi M, Zhang M (2017) A random-access algorithm based on statistics waiting in LTE-M system. In: *Proceedings of international conference on computer science and education*, IEEE, Houston, pp 214–218

Media Access Control Protocol for Cognitive Radio Networks

- [MAC in Cognitive Radio Networks](#)

Medical Internet of Things

- [Internet of Medical Things](#)

Medium Access Control

- [Interference-Aware Distributed Resource Allocation](#)

Medium Access Control in Millimeter-Wave Wireless Communications

- [Millimeter Wave MAC Layer](#)

Message-Driven Frequency Hopping

- [Index Modulation for OFDM](#)

Middleware for Wireless Sensor Networks

Flavia C. Delicato and Paulo F. Pires
DCC-IM/PESC/COPPE – Federal University of Rio de Janeiro, Cidade Universitária, Rio de Janeiro, Brazil

Synonyms

[Software platforms for WSN](#); [WSN frameworks](#); [WSN management](#)

Definitions

Wireless Sensor Networks (WSNs) are composed of the interconnection, often through wireless links, of devices equipped with sensing, processing, storage, and communication capabilities. Such devices are called sensor nodes, and a WSN can contain tens to thousands of them. Sensor nodes are tiny in size, have reduced computational resources, and are usually battery-powered, thus making energy efficiency a crucial issue to extend the WSN lifetime. Each node contains one or more sensing units, capable of performing measurements of physical variables such as vibrations, acceleration, pressure, temperature, light, among others, converting them to digital signals. The sensors act in a collaborative way, extracting environmental data, performing simple processing, and transmitting them to one or more exit points, called sink nodes, gateways or base stations, to be analyzed and further processed. The data measured by the sensors and transformed into digital signals are translated into information about a phenomenon of interest to humans.

The term middleware, although very popular in computer science, can be a bit confusing to many people. One reason for this is the myriad of different types of middleware systems in use in the everyday software development tasks. In distributed systems, middleware is usually defined as an infrastructure software that lies between the underlying operating system, networks, and hardware and the applications running on each node of the system. In general, middleware is supposed to hide the lower-level hardware and internal operation and heterogeneity of the underlying system, providing standard interfaces, abstractions, and a set of services. Another role of middleware is to provide reusable set of services that can be combined and customized to create distributed systems faster and more reliably by integrating components that may be developed by different software infrastructure suppliers (Schantz and Schmidt 2001). By hiding the inherent complexity of distribution, the use of middleware in traditional distributed computing facilitates the work of application developers.

Tasks such as concurrency control, transactions, data replication, security, and other infrastructure services are examples of services performed by a middleware. In the context of WSN, the main purposes of a middleware are supporting the development and execution of sensing-based applications and managing the utilization of the network and sensor node's resources while meeting applications' requests.

Historical Background

Despite the well-known advantages of using middleware platforms in traditional distributed systems, only from the early 2000s the researchers began to consider its adoption in the WSN design (Bonnet et al. 2001; Capra et al. 2003; Delicato et al. 2003; Murphy et al. 2001; Shen et al. 2001). Considering the simplicity of the first applications and the application-specific nature of early WSNs, the overhead of adding a middleware layer in the network design was often not worth. Therefore, the first generation of WSNs was traditionally designed from scratch in a domain-specific and task-oriented fashion. WSN systems were built as monolithic software for a specific target platform and operating system and addressing the requirements of a single target application, with little or no possibility of reusing them for newer applications. However, to accommodate the new scenarios of shared and Internet-scale WSN systems, novel systematic design approaches based on high level and preferable standardized abstractions and services were required. Thus, the support of middleware for WSNs became a crucial need to provide: (i) appropriate **system abstractions**, so that application developers can focus on the application logic and not dealing with lower level implementation details, (ii) **standardized and reusable services**, so that developers can deploy and execute application without worrying about complex, error-prone, and tedious infrastructure-level functions, such as network and resource management, context-aware adaptation, among others, (iii) **runtime environment** able to manage the execution of

multiple applications. Middleware support is also necessary to provide interoperability between different WSN systems, with external networks, as the Internet, or with enterprise systems.

To meet such demands, from the 2000s several proposals of middleware systems specially tailored for WSN emerged, each with different goals and approaches. It can be said that database-based approaches (Bonnet et al. 2001; Madden et al. 2005; Shen et al. 2001) were the first attempt to propose a rudimentary middleware layer for the project of WSNs. In the COUGAR system (Bonnet et al. 2001), the processing capabilities of sensor nodes are used to perform part of the processing of queries within the network, instead of centralizing such processing only at the sink nodes. In Bonnet et al. (2000), an SQL-based declarative query language is proposed for users to submit their queries to the WSN. Also following this line, in Shen et al. (2001), an architecture for sensor networks called SINA is presented, which provides mechanisms for query, monitoring, and submission of tasks. SINA plays the role of a middleware that abstracts WSN nodes of an RSSF as a collection of distributed objects. Users access and extract information from a WSN either by issuing declarative queries or requesting tasks using programming scripts. SINA adopts the proprietary SCTL procedural scripting language as the interface between applications and the SINA middleware. An event-driven approach is adopted as a programming model.

From these early initiatives, most of which adopted approaches with little flexibility and generally tied to specific communication protocols, a myriad of new proposals for WSN middleware began to emerge, highly diversified regarding their design approaches, abstractions and programming models. Examples of approaches are (i) event-driven, (ii) service-oriented (Mohamed and Al-Jaroodi 2011), (iii) virtual machine-based, (iv) agent-based (Chen et al. 2006; Fok et al. 2005), (v) tuple-spaces (Costa et al. 2006), (vi) component-based, and (vii) application-specific (Heinzelman et al. 2004), with the first three being the most prominent ones. Some middleware platforms use a combination of approaches.

For example, many service-oriented middleware also employs VMs in their design and development.

Virtual Machine-based approach allows developers to write application code in separate, small modules that are injected and distributed throughout the network using specialized algorithms. The goal is minimizing the overall energy consumption and resource usage. The Virtual Machine (VM) in each node then interprets the injected modules. Examples of this approach include Maté (Levis and Culler 2002), ASVM (Levis et al. 2005), and DAViM (Michiels et al. 2006).

Another widely adopted programming approach to WSN middleware is based on the concept of events (Boonma and Suzuki 2012; Silva et al. 2014; Souto et al. 2004). In such approach, components, applications, and all the other participants interact through events. Applications specify their interest in given changes of state in the monitored environment (basic events). Upon detecting an event, a sensor node sends an event notification towards interested applications. The application can also specify patterns of events (composite events). Events are propagated from the event producers (sensor nodes), to the event consumers (applications).

Among the existing approaches, the service-oriented approach has been highlighted by providing innumerable advantages in the context of WSN middleware. The service-oriented design paradigm builds software or applications as services. Service-oriented computing (SOC) is based on the service-oriented architecture (SOA) and has been traditionally used in corporate information systems. The features of SOC, such as technology neutrality, loose coupling, service reusability, composability, and discoverability, are also potentially beneficial to WSN systems. Among its advantages, it can be mentioned that this approach offers a generic and flexible programming model that favors the interoperability between different applications and the network. By exposing the functionalities of the sensors as services, it offers a more flexible architecture in comparison, for example, with

databases approaches. SQL is basically a query language, not having a suitable semantics to represent the submission of tasks and actuation activities in the WSN.

Finally, Component-Based Software Engineering (CBSE) is a modern methodology that proposes software construction by plugging software components. Based on component interoperability, this programming approach aims at building more flexible and adaptable software. Recently, some middleware proposals based on CBSE have emerged in the WSN field (Costa et al. 2007; Man et al. 2016).

Foundations

WSN technology has made tremendous progress in the last decade, drawing the attention of the scientific community and industry. WSNs are the key components of the emerging Internet-of-Things (IoT) paradigm (Man et al. 2016). With their ability to instrument the physical environment, WSNs naturally constitute the core infrastructure from which IoT systems can be built for a myriad of domains. With the introduction of the IoT, it is envisioned that future WSN deployments will have to support multiple applications simultaneously. To enable this scenario, WSNs will have to evolve from application-specific systems to networks capable of sharing data and resources among several applications, in an efficient, fair, and transparent way. Of course, such sharing creates new challenges in promoting the interoperability of different sensor nodes and WSNs while achieving the desired energy efficiency and satisfying the requirements of multiple applications. Middleware is a technology that can potentially enable WSN sharing and provide solutions to these challenges.

In short, a middleware for WSN must provide support for the development, management, deployment, and execution of sensing-and-actuation-based application. Among other functions, the middleware can decide the best protocols to be used according to the application requirements, coordinate the operation of sensors

to achieve the application goals, and intelligently manage the use of device and network resources. To efficiently provide the quality of service required by applications, it is often necessary to interact with the lower levels of protocols or even with hardware components. Middleware services can be developed to perform this interaction on behalf of applications. The provision of all these functions by a middleware system facilitates the tasks of both the application developers and of the network managers.

Since WSNs have several distinct features and constraints, conventional middleware technologies are not suitable for these networks. Therefore, the development of new middleware systems, specifically tailored to WSN is required. The following paragraphs present a set of requirements that WSN domain poses on middleware system development.

Resource Constraint: WSN nodes are constrained devices regarding their processing power, memory capacity, bandwidth, and energy. The processing and memory constraints mean that any software solution for WSNs must be computationally light. Any protocol or algorithm for such networks needs to be designed aiming at the efficient use of available resources. The execution of application tasks must be performed in a way that balances the usage of resources and the meeting of QoS requirements. Moreover, since the battery life is limited, and the exchange or even recharge of batteries in a large-scale network is undesirable or even unfeasible, a key requirement of WSN is to have mechanisms to manage the energy consumption in sensor nodes in an intelligent way.

Scale: A single WSN may consist of hundreds to thousands of nodes and to efficiently manage this kind of computational environment is challenging. The traditional concept of network management captures methods and tools related to the operation, administration, maintenance, and provisioning of networked systems. However, to manage WSN systems, composed of heterogeneous nodes with different requirements and operational properties, a paradigm shift is necessary. Upper layers need to efficiently capture dynamic system changes, and lower layers need to transform that information into

appropriate, autonomous, and scalable actions. In recent years, several extensions have been proposed for the traditional definition of systems and network management, which are specifically designed to address the increasing complexity and dynamism of these environments. Besides controlling the network operation, the high number of nodes will require mechanism to coordinate them in the execution of the required sensing tasks.

Heterogeneity: The WSN environment is composed of devices, communication links, protocols, and software components that are heterogeneous in terms of capabilities, programming abstractions, models, and development tools. Such heterogeneous nature of WSNs emerges not only from differences in capacity and features but also for other reasons including multivendor products and application requirements. Such heterogeneous environment with many sensors of different hardware platforms running applications developed by different teams will require using an abstraction layer that provides a common way to program, access the networks, and extract the sensing data.

Data Processing: Sensors generate a huge amount of data, typically consisting of time-series values, which are sampled over a specific period of time, thus characterizing a data stream. The input rate of a data stream ranges from a few bytes per second to a few gigabits per second. Such rate can be irregular, unpredictable, and bursty in nature. In addition, sensor data are subject to different degrees of accuracy. The network, by its very nature, has only the ability to sample physical processes, which in turn implies that the obtained result approximates the real parameters observed. Therefore, the data in WSNs have an extra dimension, namely, the accuracy. In this context, accuracy can be defined as the degree to which the value provided by the sensors approaches the “real” value of the monitored physical phenomenon. Some factors that contribute to the accuracy of the data measured by a sensor are its nominal accuracy (from the manufacturer) and its distance from the phenomenon, in addition to environmental noise.

Diverse application: WSN-based systems can offer its services to many applications in numer-

ous domains and environments. Different applications are likely to need different deployment architectures (e.g., event-driven and time-driven), have different requirements, and often must be served with different priorities. For instance, time critical and mission critical applications demand some level of priority in the access to the shared resources, in detriment of non-critical applications.

Context-awareness: Context is key in WSNs and their applications. A large number of sensors will generate large amounts of data, which will not have any value unless they are analyzed, interpreted, and understood. Context-aware computing stores context information related to sensor data, easing its interpretation. More specifically, location or spatial information about sensors is critical, as location plays a vital role in context-aware computing. In a large-scale WSN, interactions are highly dependent on their locations, their surroundings, and presence of other entities (e.g., other systems, neighboring nodes, and people).

Dynamicity: WSNs are subject to unpredictable changes in their execution context. Besides the intrinsic variations in the network connectivity, typical of wireless links, the availability of sensors in the network is also highly variable due to node failures, energy depletion and mobility. The network topology is also frequently changed by topology control protocols that decide when and which sensors should be turned off to save power. Within such an ad hoc environment, where there is limited or no connection to a fixed infrastructure, it will be difficult to maintain a stable network to support many application scenarios. Nodes will need to cooperate to keep the network connected and active. Moreover, the required sensing tasks, that dictate the network behavior, can also change since the emergent WSN are assumed to be used by different applications (concurrently or not). These characteristics demonstrate the high degree of dynamism in the execution context of WSNs. Therefore, it is crucial that such networks are endowed with mechanisms that provide context-awareness and adaptation.

Data-Centricity: Sensor nodes are used to collect data from its surroundings. Since

applications are interested in the data collected by a sensor and express their interests in terms of type of data and QoS requirements, the scheme for node addressing in a WSN should be based on attributes that describe the node capabilities to provide data, instead of using a scheme based on any global and unique network address. Moreover, individual readings of a node are often not relevant to the application, which is more interested in the merged or synthesized information gathered from different nodes. Considering the typical spatial density in a WSN, there is often a redundancy in the data collected for different nodes inside a given region of interest. Since communication is in general more energy consuming than data processing, individual measurements should be aggregated as near to the data source as possible so that only the resulting, relevant information is transmitted to sink nodes. Therefore, in-network, data-oriented processing should be performed in every node of the WSN, aiming at reducing data redundancy, increasing the relevance of the information reported back to the application and saving energy.

Security Issues: Since WSN can be deployed for sensitive applications like military surveillance and forecasting systems, security requirements such as confidentiality, authentication, integrity, freshness, and availability should be considered. Since most existing algorithms and security models are not suitable for WSNs, solutions specifically tailored for these networks must be designed that consider the node resource constraints.

WSN Middleware Services

All services provided by the middleware must respect the intrinsic characteristics of WSNs, especially scalability and energy efficiency. The middleware should provide robustness/fault tolerance and adaptability features to deal with the dynamic execution context of WSNs and the intermittence of wireless connections. Given the data-driven nature of these networks, middleware must provide data-centric mechanisms for processing and querying data within the network. In addition, application-specific knowledge can

be used to optimize network operation, while sharing resources across multiple concurrent applications. Due to the restricted capabilities of nodes, the middleware must be light in terms of its communication and processing demands. The main services to be provided by a WSN middleware are described as follows.

Resource Management: WSN middleware should provide services to manage and optimize the network resource usage in an efficient way while meeting the application requirements. One example of such services is to select the set of nodes that will participate in a given sensing task, always considering the tradeoff between meeting the application requirements and optimizing energy consumption. Another useful service is to implement data fusion techniques to merge sensor readings of individual sensors into a high-level result to be sent to the application, thus saving transmission energy while increasing the data accuracy.

Context Management and Dynamic Adaptation: Considering the high dynamism of WSN environments, in which devices may become unavailable for a variety of reasons, middleware platforms must provide strategies for dynamic adaptation, thus ensuring the availability and quality of the applications during their execution. This is particularly important for applications in critical domains, since flaws or degradation of quality parameters in such applications can be a threat to human life and health. Thus, WSN middleware must remain available and functioning properly in this dynamic environment. To do so, it needs to provide mechanisms for collecting and analyzing contextual data, and responding to changes in the execution context, without decreasing the WSN system overall performance. Adaptive measures must be applied whenever the target operational performance metrics are not met. Examples of adaptive actions are turning on/off sensors, changing the data sensing/sending rates, enabling/disabling data aggregation functions, and changing the routing protocol in use on the network. QoS metrics and adaptation plans are managed by the middleware via policies and rules predefined by applications.

Resource and Service Discovery: WSN resources are very heterogeneous, including from low-level processing, communication, and sensing capabilities of individual sensor nodes to high-level services such as processed data and complex events detected from analyses of raw sensing data. For applications to make use of these resources, middleware must provide mechanisms for their discovery. In addition, in the absence of network infrastructure (since WSN are often ad hoc in nature), the middleware must provide two levels of service discovery. The internal level of discovery is used so that the sink nodes are aware of all the capacities of the sensor nodes that make up the network. Likewise, neighbors must know each other's resources. This is done by advertising mechanism that may rely on protocols and data formats common within a same WSN. The external level of service discovery is used by an application to find out which WSN provide the services it requires and how to access those services. In this case, the middleware should provide a standardized interface to access the WSN resources, abstracting its specific formats and protocols. Considering the dynamic environment of WSNs, assumptions related to global and deterministic knowledge of the resources' availability do not hold. Moreover, considering the high scale and geographical features of the WSN deployment, human intervention for resource discovery is infeasible. Therefore, an important requirement for WSN middleware is providing automatic resource discovery, based on high-level descriptions of application functional and QoS requirements.

Code management: Deploying code in a WSN environment is challenging and should be directly supported by the middleware (Razzaque et al. 2016). In particular, code allocation and code migration services are required. Code allocation selects the set of devices or sensor nodes to be used to accomplish a user or application-level task. Code migration transfers one node's code to another, potentially reprogramming nodes in the network, for instance to address requirements of arriving applications in a dynamic way. Approaches such

as VM and agent-based help provide this type of service, since they offer a common execution environment to run migrated code.

In addition to these, there are several generic services that a WSN middleware can provide, such as: naming, location, security, and time synchronization.

Key Applications

Military applications such as target detection, battlefield surveillance, and counterterrorism originally motivated WSN applications. However, sensor nodes are very flexible and can be used for continuous data sensing, event detection, and local actuator control, as well as other tasks. The advantages of WSN over traditional networks resulted in many other potential applications, ranging from infrastructure security to industrial/factory control, environment and habitat monitoring, healthcare applications, home automation, and traffic control. The design and development of a successful middleware must address challenges posed by WSN features on one hand and the application demands on the other hand.

Cross-References

- [Data Gathering in Wireless Sensor Networks](#)
- [Information-Centric Wireless Sensor Networks](#)
- [Protocol Stack Architecture for Wireless Sensor Networks](#)

References

- Bonnet P, Gehrke J, Seshadri P (2000) Querying the physical world. *IEEE Pers Commun* 7:10–15
- Bonnet P, Gehrke J, Seshadri P (2001) Towards sensor database systems. In: *International conference on mobile data management*. Springer, London, pp 3–14
- Boonma P, Suzuki J (2012) Chapter 41, TinyDDS: an interoperable and configurable publish/subscribe middleware for wireless sensor networks. In: *Information Resources Management Association (ed) Wireless technologies: concepts, methodologies, tools and applications*. IGI Global, Hershey, pp 819–846

- Capra L, Emmerich W, Mascolo C (2003) Carisma: context-aware reflective middleware system for mobile applications. *IEEE Trans Softw Eng* 29:929–945
- Chen M, Kwon T, Yuan Y, Leung VC (2006) Mobile agent based wireless sensor networks. *J Comput* 1:14–21
- Costa P, Mottola L, Murphy AL, Picco GP (2006) TeenyLIME: transiently shared tuple space middleware for wireless sensor networks. In: *Proceedings of the international workshop on middleware for sensor networks*. ACM, New York, pp 43–48
- Costa P et al (2007) The RUNES middleware for networked embedded systems and its application in a disaster management scenario. In: *Fifth annual IEEE international conference on pervasive computing and communications, PerCom '07*. IEEE, White Plains, pp 69–78
- Delicato FC, Pires PF, Pirmez L, da Costa Carmo LFR (2003) A flexible middleware system for wireless sensor networks. In: Endler M (ed) *Proceedings of the ACM/IFIP/USENIX 2003 international conference on middleware (Middleware '03)*. Springer, New York, pp 474–492
- Fok C-L, Roman G-C, Lu C (2005) Rapid development and flexible deployment of adaptive wireless sensor network applications. In: *Proceedings of 25th IEEE international conference on distributed computing systems, ICDCS 2005*. IEEE, Columbus, pp 653–662
- Heinzelman WB, Murphy AL, Carvalho HS, Perillo MA (2004) Middleware to support sensor network applications. *IEEE Netw* 18:6–14
- Levis P, Culler D (2002) Maté: a tiny virtual machine for sensor networks. *ACM SIGPLAN Not* 10:85–95. ACM
- Levis P, Gay D, Culler D (2005) Active sensor networks. In: *Proceedings of the 2nd conference on symposium on networked systems design & implementation (NSDI '05)*, vol 2. USENIX Association, Berkeley, pp 343–356
- Madden SR, Franklin MJ, Hellerstein JM, Hong W (2005) TinyDB: an acquisitional query processing system for sensor networks. *ACM Trans Database Syst (TODS)* 30:122–173
- Man KL, Hughes D, Guan S-U, Wong PW (2016) Middleware support for dynamic sensing applications. In: *2016 International conference on platform technology and service (PlatCon)*. IEEE, Jeju, pp 1–4
- Michiels S, Horré W, Joosen W, Verbaeten P (2006) DAVIM: a dynamically adaptable virtual machine for sensor networks. In: *Proceedings of the international workshop on middleware for sensor networks (Mid-Sens '06)*. ACM, New York, pp 7–12
- Mohamed N, Al-Jaroodi J (2011) A survey on service-oriented middleware for wireless sensor networks. *SOCA* 5:71–85
- Murphy AL, Picco GP, Roman G-C (2001) Lime: a middleware for physical and logical mobility. In: *21st International conference on distributed computing systems*. IEEE, Mesa, pp 524–533
- Razzaque MA, Milojevic-Jevric M, Palade A, Clarke S (2016) Middleware for internet of things: a survey. *IEEE Internet Things J* 3:70–95
- Schantz RE, Schmidt DC (2001) Middleware for distributed systems: evolving the common structure for network-centric applications. In: *Encyclopedia of software engineering*, vol 1. Wiley, New York, pp 1–9
- Shen C-C, Srisathapornphat C, Jaikao C (2001) Sensor information networking architecture and applications. *IEEE Pers Commun* 8:52–59
- Silva JR, Delicato FC, Pirmez L, Pires PF, Portocarrero JM, Rodrigues TC, Batista TV (2014) PRISMA: a publish-subscribe and resource-oriented middleware for wireless sensor networks. In: *Proceedings of the tenth advanced international conference on Telecommunications*. Citeseer, Paris, p 8797
- Souto E, Guimarães G, Vasconcelos G, Vieira M, Rosa N, Ferraz C (2004) A message-oriented middleware for sensor networks. In: *Proceedings of the 2nd workshop on middleware for pervasive and ad-hoc computing (MPAC '04)*. ACM, New York, pp 127–134

Millimeter Wave Beam Alignment and Adjustment

► Millimeter Wave Beam Training and Tracking

Millimeter Wave Beam Training and Tracking

Yongming Huang¹, Jianjun Zhang⁴, Chen Zhang^{2,3,4}, and Ming Xiao⁵

¹School of Information Science and Engineering, Southeast University, Nanjing, China

²Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

³The George Washington University, Washington, DC, USA

⁴Southeast University, Nanjing, China

⁵Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

[Millimeter Wave Beam Alignment and Adjustment](#)

Definition

Beam training is the search procedure that the transmitter and/or the receiver selects one or multiple beams from its beam training codebook (a set of predefined beams), so as to achieve the necessary link budget for subsequent communications via transmitting or receiving signals with these beams. Beam tracking is the adjustment procedure that the transmitter and/or the receiver adjusts its beams to guarantee the link stability for time-varying propagation environment. Beam training and tracking are usually used in the millimeter wave communications to combat the severe path-loss of signal propagation and the user mobility, respectively.

Historical Background

Millimeter wave (mmwave) communications operating in the band of 30–300 GHz has attracted much attention recently and also been recognized as a promising technology for future mobile networks, owing to its abundant spectrum resources (Xiao et al. 2017). However, the propagation of mmwave signal usually suffers a large path-loss, which severely limits the coverage range of mmwave communications. To address this issue, large antenna arrays are usually equipped in the mmwave communication system, so that the array gain can be exploited to combat the path-loss and increase the communication distance (Heath et al. 2016). Thanks to the short wave length of mmwave signals, it is possible to pack a large number of antennas into a compact size.

Due to the large path-loss of mmwave signals, the necessary link budget usually cannot be satisfied before achieving the array gain provided by the large antenna arrays. Therefore, at the beginning of communications, it is essential for the transmitter and the receiver to determine their respective beams, so as to achieve the array gain and build reliable communications links (Zhang et al. 2017a). To achieve this goal, beam training is designed to select the preferred transmit beam and/or the preferred receive beam from their

respective beam training codebook, i.e., a set of predefined beams. Beam training plays a key role in mmwave communications due to its importance in building reliable mmwave communication links, acquiring channel state information (CSI), achieving desired rate of data transmission, and enlarging the cell coverage area.

As a viable approach, beam training was adopted in mmwave multi-gigabit wireless local area network (WLAN) (Tsang et al. 2011) and wireless personal area network (WPAN) (Wang et al. 2009), where analog antenna array architecture and analog beams were considered. Compared with the straightforward method, i.e., to exhaustively train all possible beams in the codebook, then to find the best one or several beams for transmission, a hierarchical-search-based beam training scheme was proposed in Hur et al. (2013) to reduce the training overhead based on multi-resolution or variant beamwidth analog beams. Specifically, all possible wide beams are first trained and the best is selected. Then the beams with narrower beamwidth are trained within the coverage range of this selected best wide beam and the best is chosen. This procedure is repeated until the optimal beam with acceptable beamwidth is found. Later, to improve the accuracy of the hierarchical-search-based training scheme while still maintain its low training overhead, a beam training codebook for the analog architecture was devised in Xiao et al. (2016) by exploiting subarray and deactivation antenna processing techniques. To address the problem of error propagation in hierarchical-search-based training schemes, a novel design of training codebook was proposed in Zhang et al. (2017a) based on the hybrid antenna array architecture, which can provide more flexibility in designing training beams compared to the analog architecture (Alkhateeb et al. 2014). To further reduce the training overhead, a beam training approach was proposed in (Kokshoorn et al. 2017), based on overlapped beam patterns. The beam training approaches aforementioned are mainly designed for single data stream transmission. To enable multiple data stream transmission, a cooperative multi-subarray beam training approach was proposed in Zhang

et al. (2017b) recently for a codebook-based beamforming system.

All the abovementioned beam training schemes correspond to decoupled transmission design (called non-interleaved training) since the complete effective CSI should be obtained first for the full or selected partial beam codebook. When the size of the training beam codebook is comparative to the BS antenna number for satisfactory performance, heavy training overhead is unavoidable. The decoupled nature of non-interleaved schemes thus imposes limitations on the trade-off between training overhead and performance. In Zhang et al. (2018), an interleaved training design was proposed to achieve favorable trade-off between the outage performance and the training overhead where the BS and/or users monitors the training procedure to avoid training redundancy, i.e., to find just enough beams to avoid an outage; thus, the training length depends on the channel realization, and the termination of the training process depends on previous training results, leading to smaller average training overhead.

To achieve a large array gain, the beamwidth of the training beams shall be narrow, and thus the size of the beam training codebook could be large, which leads to a significant beam training overhead. This problem becomes even more serious for the mobile mmwave scenario, since there may be no enough time to perform beam training repeatedly. To address this issue, it is necessary to resort to beam tracking, which is realized by appropriately adjusting the beam directions. Beam tracking has been studied for a long time. The first beam tracking approach in mmwave communication (Wang et al. 2009; Kim and Lee 2015) also finds the candidate beams besides the desired ones in each training and then checks the feasibility of candidate beams in the next training. If some time-varying models are assumed, other beam training approaches can be used for beam tracking as well. Recently, beam tracking solutions based on conventional Kalman filter have drawn considerable attention, by exploiting the temporal correlation of time-varying channels (Jayaprakasam et al. 2017; Zhang et al. 2016; Va et al. 2016).

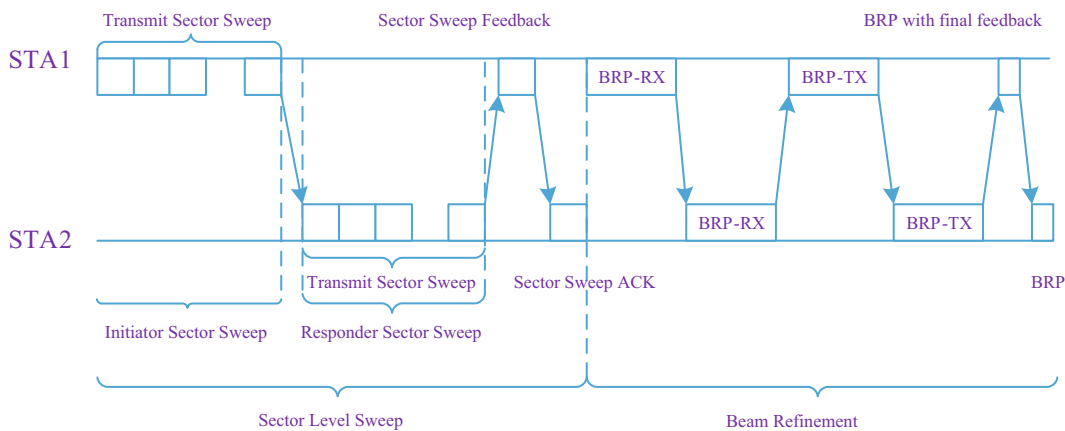
Foundations

Beam training is a viable approach to achieve the goals of building reliable mmwave communication links, enlarging the coverage area and so on in mmwave communications. In fact, beam training has been widely used in practice and adopted by some communication standards, e.g., IEEE 802.11ad and IEEE 802.15.3c. To better illustrate the principle of beam training and tracking, we take the scheme of single-beam training in IEEE 802.11ad as an example.

Beam training, termed as beamforming (BF) training in the IEEE 802.11ad, is a mechanism used by a pair of stations (STAs), e.g., one transmitter and one receiver, to achieve the necessary link budget for subsequent communication. It consists of three phases, i.e., sector-level sweep (SLS), beam refinement protocol (BRP), and beam tracking. An example of BF training procedure is provided in Fig. 1 for clarity. The STA that initiates BF training is referred to as the initiator, and the recipient STA that participates in BF training with the initiator is referred to as the responder. BF training starts with a sector-level sweep (SLS) from the initiator. The purpose of the SLS phase is to enable communications between the two participating STAs. To accelerate the speed of beam training and further reduce the training overhead, in the stage of SLS, the beams with wide beamwidth are considered at first, as shown in Fig. 2a. The SLS procedure of the BF training is presented below in details.

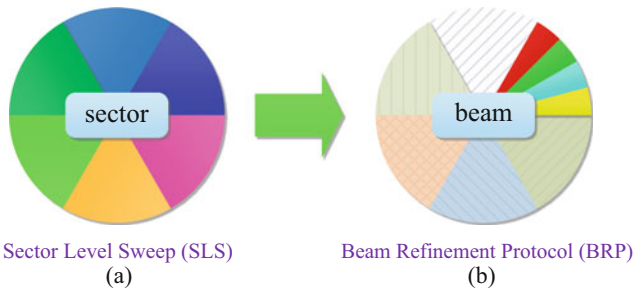
Sub-phase 1: In this sub-phase of SLS, the initiator uses sector-wide beams in turn to sweep its beam space, while the responder receives the transmitted signals using omnidirectional or quasi-omnidirectional beam, as shown in Fig. 3. Based on the received signals, the responder can determine the best sectorized transmit beam of the initiator. Moreover, the index of the best sectorized transmit beam should be fed back to the initiator.

Sub-phase 2: As shown in Fig. 4, sector-wide beams are used in turn by the responder to sweep its beam space, while omnidirectional or quasi-omnidirectional beam is used by the initiator to receive signals. Based on the received signals,

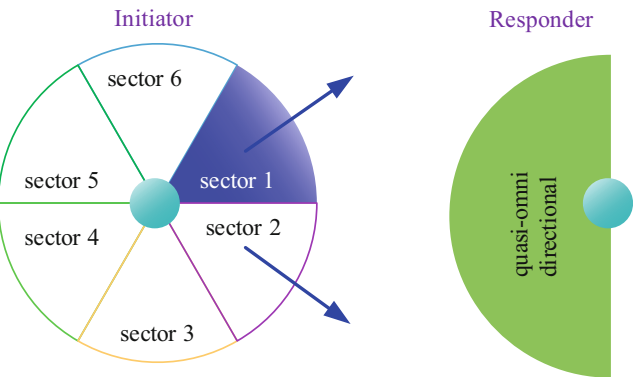


Millimeter Wave Beam Training and Tracking, Fig. 1 An example of beamforming training in IEEE 802.11ad

Millimeter Wave Beam Training and Tracking, Fig. 2 Beam training in IEEE 802.11ad: (a) sector-level sweep; (b) beam refinement protocol



Millimeter Wave Beam Training and Tracking, Fig. 3 Beam training in IEEE 802.11ad – sub-phase 1 of SLS



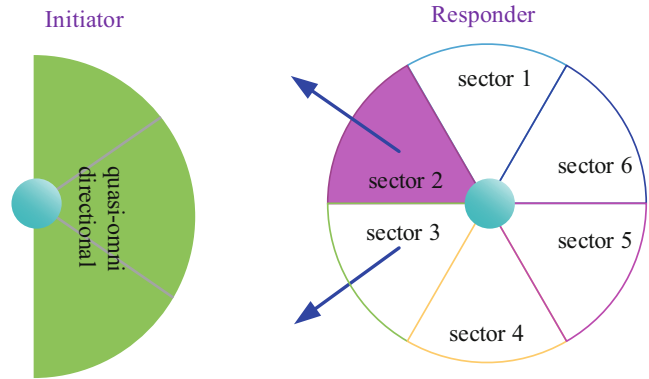
the initiator can obtain its best sector transmit beam. Moreover, the initiator can determine the best sector transmit beam of the responder and should feedback the index to the responder.

Sub-phase 3: The initiator uses the best sector transmit beam obtained in sub-phase 2 to transmit signals, as shown in Fig. 5. The responder uses

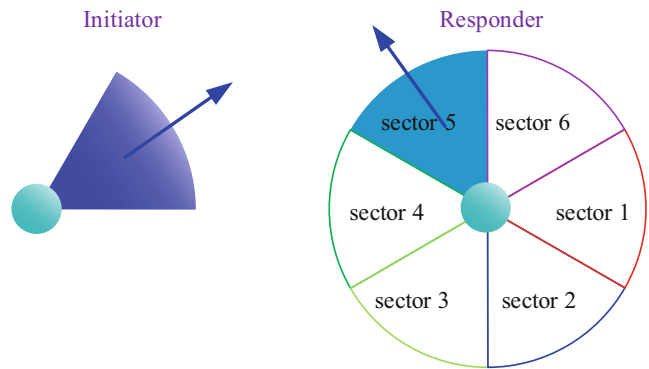
sector-wide beams in turn to receive signals. Based on the received signals, the responder can obtain its best sector transmit beam and also determine its best sector receive beam.

Sub-phase 4: In this sub-phase, the responder uses its best sector transmit beam to transmit signals, while the initiator uses sector-wide beams

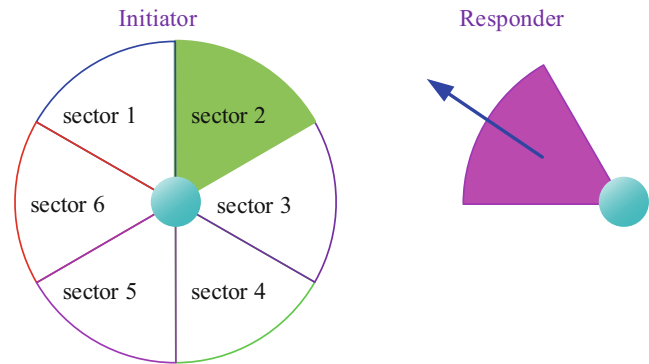
Millimeter Wave Beam Training and Tracking,
Fig. 4 Beam training in
 IEEE 802.11ad –
 sub-phase 2 of SLS



Millimeter Wave Beam Training and Tracking,
Fig. 5 Beam training in
 IEEE 802.11ad –
 sub-phase 3 of SLS



Millimeter Wave Beam Training and Tracking,
Fig. 6 Beam training in
 IEEE 802.11ad –
 sub-phase 4 of SLS



to receive signals, as shown in Fig. 6. Based on the received signals, the initiator can determine its best sector receive beam.

When the four sub-phases of SLS aforementioned are finished, both the initiator and the responder can determine their best sector transmit and receive beams. Next, a beam refinement protocol (BRP) may follow, if requested by either the initiator or the responder. The purpose of the

BRP phase is to enable iterative refinement of the antenna weight vector of both transmitter and receiver at both participating STAs. In the phase of BRP, the beams with narrow beamwidth are considered, as shown in Fig. 2b. The training procedure in this phase is similar to that in the SLS.

Beam tracking can effectively reduce the training overhead, especially for mmwave mobile

environment. Similarly, a relatively simple beam tracking approach, which is a simplified and improved version of the beam training approach in (Wang et al. 2009), is used to show the basic principle and operational procedure. Due to its quite low complexity and easy implementation, the beam tracking approach has been adopted in current mmwave communication standard, e.g., IEEE 802.11ad. The key idea is to select several candidate beams (or beam pairs) instead of only the optimal beam (or beam pair), during the beam training procedure. For example, when the beam pair with the highest SNR is selected, the beam pairs having the second and third highest SNRs are also retained. When the channel varies, only the candidate beam pairs are tested and switched on to keep the SNR above a certain threshold. The complete beam training procedure is performed, only when all the candidate beam pairs fail. In IEEE 802.11ad, the beam refinement is performed periodically to track beam directions in the third phase.

It should be pointed out that in the literature, there also exist other beam training schemes, which can further improve the performance of beam training. However, most of these beam training schemes are extended from the schemes standardized in IEEE 802.11ad or IEEE 802.15.3c. Moreover, these beam training schemes require carefully designed training codebooks. We would like to recommend readers to refer to Zhang et al. (2017a), Tsang et al. (2011), Wang et al. (2009), Hur et al. (2013), Xiao et al. (2016), Alkhateeb et al. (2014), and Kokshoorn et al. (2017) for details.

The beam training and tracking approaches aforementioned are mainly applied to the case of single data stream. It is very desired to transmit multiple data streams simultaneously in practice. In this case, if conventional beam training schemes are still used, the resulting complexity and training overhead may be prohibitive. To address this issue, the idea of multi-subarray cooperation was introduced in Zhang et al. (2017b), where a codebook-based beamforming mmwave system was considered. It is recommended to refer to Zhang et al. (2017b) for more details.

Key Applications

Millimeter wave wireless local area networks (WLAN), 802.11ad, 802.11aj, 802.15.3c; wireless display; virtual reality or enhanced reality; automatic sync applications (e.g., uploading images from a camera to a PC); millimeter wave cellular networks.

Cross-References

- [Millimeter Wave Channel Access](#)
- [Millimeter Wave Channel Modeling](#)
- [Millimeter Wave Massive MIMO](#)

References

- Alkhateeb A, El Ayach O, Leus G, Heath R (2014) Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE J Sel Top Sign Process* 8(5):831–846
- Heath RW, Gonza'lez-Prelcic N, Rangan S, Roh W, Sayeed AM (2016) An overview of signal processing techniques for millimeter wave mimo systems. *IEEE J Sel Top Sign Process* 10(3):436–453
- Hur S, Kim T, Love D, Krogmeier J, Thomas T, Ghosh A (2013) Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans Commun* 61(10):4391–4403
- Jayaprakasam S, Ma X, Choi JW, Kim S (2017) Robust beam-tracking for mmwave mobile communications. *IEEE Commun Lett* 21(12):2654–2657
- Kim J, Lee I (2015) 802.11 wlan: history and new enabling mimo techniques for next generation standards. *IEEE Commun Mag* 53(3):134–140
- Kokshoorn M, Chen H, Wang P, Li Y, Vucetic B (2017) Millimeter wave mimo channel estimation using overlapped beam patterns and rate adaptation. *IEEE Trans Signal Process* 65(3):601–616
- Tsang YM, Poon ASY, Addepalli S (2011) Coding the beams: improving beamforming training in mmwave communication system. In: 2011 IEEE global telecommunications conference – GLOBECOM 2011, pp 1–6
- Va V, Vikalo H, Heath RW (2016) Beam tracking for mobile millimeter wave communication systems. In: 2016 IEEE global conference on signal and information processing (GlobalSIP), pp 743–747
- Wang J, Lan Z, Pyo C-W, Baykas T, Sum C-S, Rahman M, Gao J, Funada R, Kojima F, Harada H, Kato S (2009) Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems. *IEEE J Sel Areas Commun* 27(8):1390–1399

- Xiao Z, He T, Xia P, Xia XG (2016) Hierarchical codebook design for beamforming training in millimeter-wave communication. *IEEE Trans Wirel Commun* 15(5):3380–3392
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis GK, Björnson E, Yang K, Chih-Lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Top Sign Process* 35(9):1909–1935
- Zhang C, Guo D, Fan P (2016) Tracking angles of departure and arrival in a mobile millimeter wave channel. In: 2016 IEEE international conference on communications (ICC), pp 1–6
- Zhang J, Huang Y, Shi Q, Wang J, Yang L (2017a) Codebook design for beam alignment in millimeter wave communication systems. *IEEE Trans Commun* 65(11):4980–4955
- Zhang J, Huang Y, Zhang C, He S, Xiao M, Yang L (2017b) Cooperative multi-subarray beam training in millimeter wave communication systems. In: GLOBE-COM 2017–2017 IEEE global communications conference, pp 1–6
- Zhang C, Jing Y, Huang Y, Yang L (2018) Interleaved training and training-based transmission design for hybrid massive antenna downlink. *IEEE J Sel Top Sign Process* 12(2):541–556

Millimeter Wave Channel Access

Chen Zhang^{1,2,3}, Yongming Huang⁴, and Yili Xia³

¹Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

²The George Washington University, Washington, DC, USA

³Southeast University, Nanjing, China

⁴School of Information Science and Engineering, Southeast University, Nanjing, China

Synonyms

Millimeter wave multiple access

Definition

Millimeter wave (mmWave) channel access is the technology that enables more than one users connected to the same system resource block for

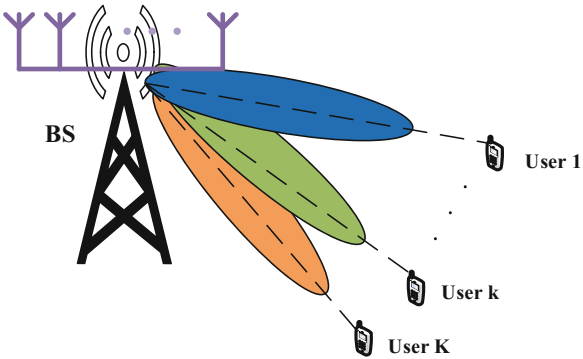
concurrent transmission in wireless communications networks operating at mmWave band.

Historical Background

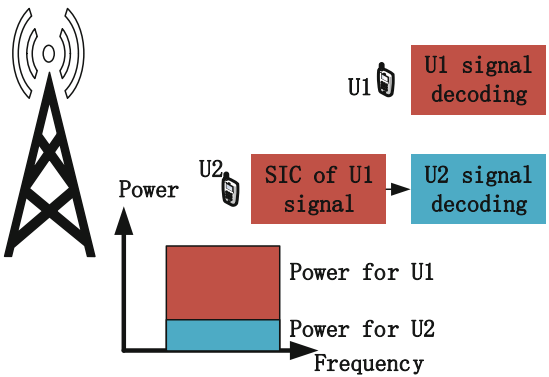
Channel-access technologies are used for supporting multiple-user wireless communication over networks. Several fundamental types of channel access schemes have been widely investigated in Sub-6 GHz microwave wireless communications. For example, in frequency division multiple access (FDMA), time division multiple access (TDMA), and code division multiple access (CDMA), multiple users utilize different frequency bands, time slots, and codes, to establish their own links with the base station (BS), respectively (Goldsmith 2005). These orthogonal multiple access (OMA) technologies have been already utilized in Sub-6 GHz communications systems such as 3G/4G mobile communications systems. They are also applicable to mmWave wireless communication systems. For example, the 802.11ad standard operating at 60 GHz mmWave spectrum adopts the TDMA Scheme (802.11ad 2012).

Compared with these OMA schemes, nonorthogonal multiple access (NOMA) schemes are less used in existing wireless communications standards. As a typical spatial-domain NOMA technology, spatial division multiple access (SDMA) was introduced in 1990s (Roy and Ottersten 1991) (Fig. 1). In SDMA, different beams are used to transmit messages of different users within the same resource block (e.g., time-frequency block), and each beam is optimized to make sure a large signal gain for its dedicated user and small gains for other users. The SDMA scheme has been already supported in the LTE standards since LTE Release 8 (Technical Specifications n.d.), and IEEE 802.11 ac is the first major wireless communications standard to integrate the SDMA scheme (Wireless LAN Medium Access Control 2011). Massive MIMO is the latest technology along this direction (Rusek et al. 2013), in which the BS is equipped with large-scale antenna arrays, e.g., with hundreds of antennas, to serve tens of users via SDMA. In massive MIMO systems operating

**Millimeter Wave
Channel Access, Fig. 1**
SDMA for K users



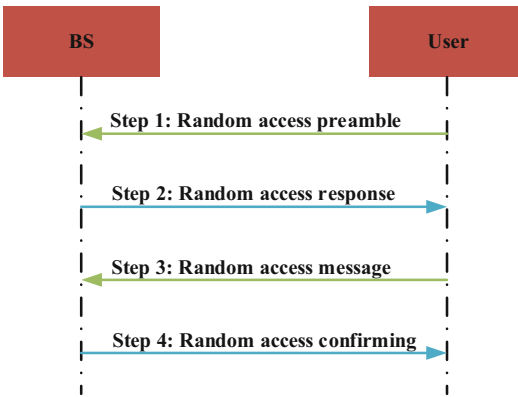
**Millimeter Wave
Channel Access, Fig. 2**
Two-user SISO-NOMA systems



at Sub-6 GHz, due to massive antennas at the BS and relatively rich scattering in the propagation, users' antenna-domain channels become less nonorthogonal. This enables high user multiplexing gains, even with simple linear precodings, e.g., maximum ratio transmit (MRT) and zero-forcing (ZF), at the BS (Zhang et al. 2017a).

Another NOMA technology under the spotlight in academic is the power-domain NOMA (usually called NOMA for short), which has been considered as one of the candidates for improving the spectral efficiency and connectivity density in 5G communications (Ding et al. 2014; Dai et al. 2015). The 3rd Generation Partnership Project (3GPP) (a collaboration between groups of telecommunications standards associations) also considers NOMA as a candidate channel-access technology in 5G communications and discussed it at RAN plenary #66 in the name of Multi-User Superposition Transmission (MUST) (<https://portal.3gpp.org/desktopmodules/Specifications/>

[SpecificationDetails.aspx?specificationId=2912](https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2912)). In conventional single-input-single-output (SISO) NOMA systems with single-antenna BS and users, different signals from different users are superimposed on each other at the BS. Multiple users share the same resource block and detect their signals via successive interference cancellation (SIC). Specifically, as shown in Fig. 2, for a two-user SISO-NOMA system, the BS serves a far-end user U1 and a near-end user U2, both users share the same resource block with a certain transmit power allocation $P_1 > P_2$. Under this scheme, user U1 decodes its own message by directly treating the message of user U2 as the background noise, while user U2 first decodes the message of user U1 and then decodes its own message after removing the message of user U1. It has been shown that the spectral efficiency can be enhanced at the cost of an increased receiver complexity compared to conventional OMA systems, and larger gaps among users' channel gains and better



Millimeter Wave Channel Access, Fig. 3 Random access procedure in Sub-6 GHz cellular networks

power allocation policy are both beneficial to the performance of NOMA systems (Zhu et al. 2017; Wang et al. 2017a).

The above-introduced channel-access technologies are mainly used in the data transmission stage. In other communication control stages, e.g., the initialization of user access and cell handover, the random access technology plays an important role. As shown in Fig. 3, four steps in the random access procedure in Sub-6 GHz cellular networks are as follows (Jeong et al. 2015): In step 1, the user transmits one selected preamble signature to the BS by using the resource block indicated in the system information broadcast by the BS. In step 2, if the BS successfully detects the user's preamble, it transmits the random-access response to the user, including the index of the detected preamble sequence, the uplink timing information, and the indication of the resource allocation for the next step. In step 3, the identity of the user is transmitted to the BS. Finally, in step 4, the identity of the user is transmitted back to the user from the BS to confirm that the random access procedure is successfully completed for the user.

Foundations

Since shorter wavelengths at the mmWave band make it easier to equip massive antennas in the same physical space, massive MIMO

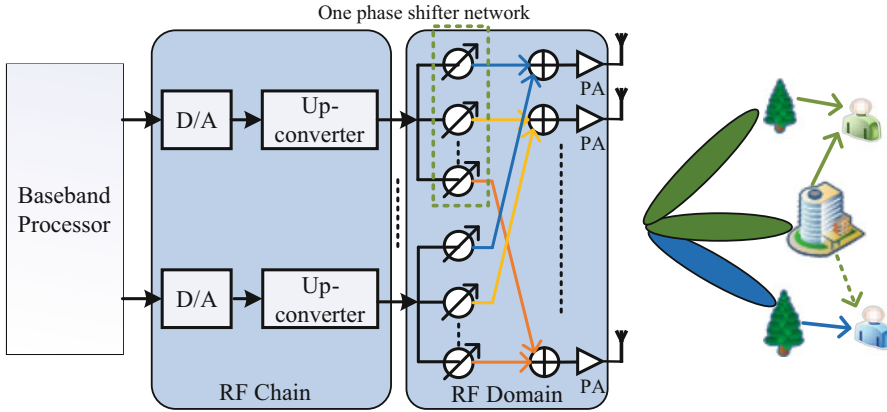
becomes the representative technology to build up mmWave communication systems. Due to the large spatial degrees of freedom in massive MIMO, the spatial resolution-based channel access technologies are of great importance. In what follows, the new features of channel and system structures in mmWave massive MIMO systems which affect the design of channel access will be first introduced. Then, the application of SDMA and its combination with NOMA and random access will be subsequently discussed for mmWave massive MIMO systems.

Features of mmWave Channel and System Structures

Compared to the Sub-6 GHz systems, in mmWave massive MIMO systems, channels become spatially sparse and highly directional due to limited scattering. This, however, results two adverse effects. On one hand, different users' antenna-domain channels become less orthogonal due to the reduced effective channel dimensions, especially for users with similar locations. On the other hand, channel powers are distributed in fewer scatterers, which indicates that using partial channel scatterers, e.g., via directional beamforming, for transmission results in smaller performance losses. In addition, the mmWave channel is typically with a large path loss, and this undesirable physic demands a large array gain. Moreover, due to their hardware costs and power consumptions, as shown in Fig. 4, one typical structure of mmWave massive MIMO systems is the hybrid analog/digital one with limited RF chains and phase shifter networks. Under this structure, fewer degrees of freedom are available for beamforming design, especially for the practical low-resolution phase shifters, leading to a coarser beamforming capability.

SDMA in mmWave Systems

Under the hybrid beamforming structure, its coarser beamforming capability results in a weaker capability of user interference suppression. Meanwhile, due to the channel nonorthogonality in antenna-domain, popularly used linearly precoding methods in Sub-6 GHz massive MIMO systems may have nonnegli-



Millimeter Wave Channel Access, Fig. 4 MmWave massive MIMO system with hybrid beamforming and limited scattering channels. D/A is digital to analog converter. Up-converter moves the frequency of the input signal to

higher frequency. One phase shifter network includes the same number of phase shifters as the number of antennas. PA is the power amplifier

gible performance degradation. Furthermore, another factor making designs of SDMA more challenge is the incompleteness of channel state information (CSI) at the BS. Specifically, due to the possible low prebeamforming signal-to-noise-ratio (SNR), massive antennas, and limited RF chains, the mmWave massive system tends to possess only partial or long-term CSI at the BS. Practically, the beamforming design should be based on long-term channel statistics (channel direction and spatial profile of average channel power) and the beamspace CSI, i.e., instantaneous effective CSI for limited analog beams (each analog beam corresponds to one phase setting of one phase shifter network). Therefore, the mmWave channel characteristics, the less flexible hybrid structure, and the CSI incompleteness at the BS independently or jointly make the transmit precoding/beamforming and receive combining (when users also have multiple antennas) more challenging (Sun et al. 2016).

Some effective hybrid beamforming algorithms have been proposed to guarantee a desirable SDMA performance. For example, in Alkhateeb et al. (2015), a decoupled baseband/analog design was proposed where by ignoring the user interference, the BS first finds the best beam for each user via beam training (Zhang et al. n.d.) and then conducts the baseband digital beamforming to reduce the residual user

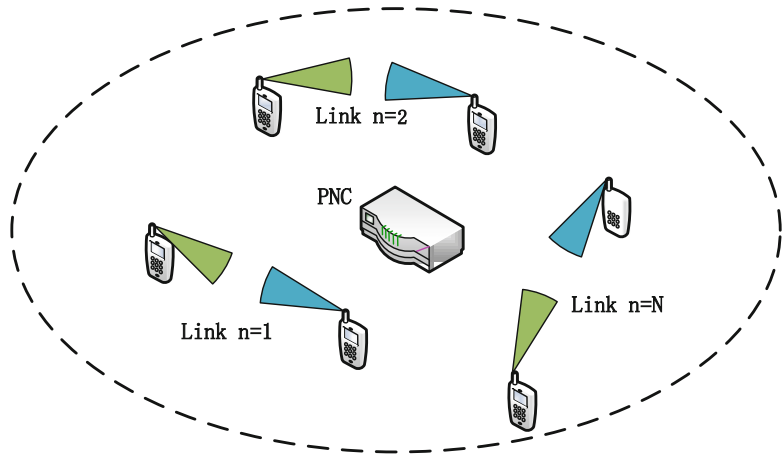
interference based on beamspace CSI. In Li et al. (2017), the capacity of beamspace channel is used as the optimization object for beamforming design, while the analog beamforming is achieved based on channel statistics only. In He et al. (2017a), the codebook (with finite analog beams)-based hybrid beamforming optimization is simplified to a joint codeword selection and beamforming design problem where both the baseband and analog designs are achieved based on complete beamspace CSI.

Spatial Sharing Multiple Access

The spatial sharing-based multiple access has been adopted into the wireless local area networks (WLAN) standard 802.11 ad (Hany and Widmer 2017). Similarly, for wireless personal area network (WPAN) standard 802.15.3c operating at mmWave bands (Baykas et al. 2011), many researchers are also interested in applying this spatial sharing idea to increase system throughput. Specifically, via utilizing the high propagation loss and antenna directionality, multiple links (each link is made up of one transmitter and one receiver) are scheduled to transmit concurrently in the same time slot. This combination of spatial sharing-/reusing-based multiple access with the original TDMA is named the spatial time division multiple access STDMA (Qiao et al. 2012) (Figs. 5).

Millimeter Wave Channel Access, Fig. 5

STDMA in WPAN. The piconet coordinator (PNC) provides the basic timing for the piconet with the beacon and manages QoS requirements, power save modes, and access control to the piconet



User Scheduling

The performance of SDMA is highly dependent on channel characteristics. For example, in typical mmWave network situations where users are in similar locations and near line-of-sight (LoS) channels exist, SDMA can only avoid large user interference via extremely fine beamforming. These mmWave channel characteristics make the user scheduling necessary, i.e., it is more practical to serve some users via SDMA and other users in orthogonal system resources. Practically, due to the aforementioned CSI incompleteness at the BS, the user scheduling should be based on the long-term channel statistics (Zhang et al. 2017b) and/or partial beamspace CSI (He et al. 2017b). Therefore, the optimal user scheduling for mmWave hybrid massive MIMO systems involves the joint design of user selection, analog beam assignment/design, and baseband digital beamforming, which is more complicated than its counterpart of Sub-6 GHz systems. In He et al. (2017b), the limited beamspace CSI-based joint analog beam selection and user scheduling task is formulated as a nonconvex and combinatorial optimization problem, and two codebook-based low-complexity methods are proposed to address this problem.

For the STDMA transmission, the optimal concurrent transmission scheduling problem can be converted to a Knapsack problem which is NP-complete (Pisinger 2005). In Sum et al. (2009), multiple communication links are scheduled within the same time slot if the

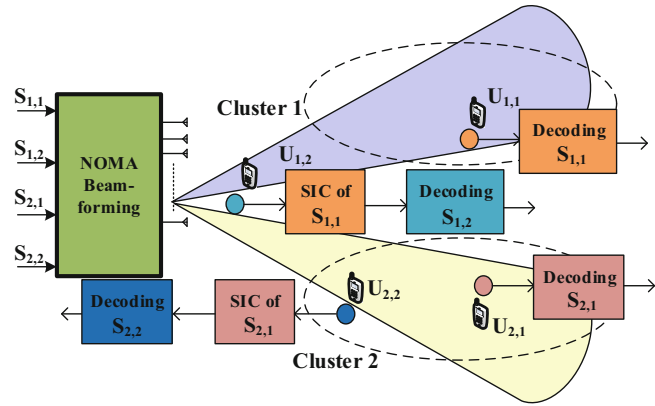
accumulated interference in this slot is below a specific threshold. In Qiao et al. (2012), the scheduling design is optimized to maximize the number of concurrent flows such that the quality of service (QoS) requirement of each flow is satisfied. This optimization problem is decomposed and solved via a heuristic scheduling algorithm.

MIMO-NOMA in mmWave Systems

With multiple antennas at the BS in mmWave systems, the combination of NOMA and MIMO becomes a natural choice. It has been shown that (1) NOMA can achieve a higher channel capacity than OMA in both mmWave uplink and downlink systems (Naqvi and Hassan 2016); and (2) Enormous capacity improvement can be achieved by mmWave MIMO-NOMA, compared to existing LTE systems (Zhang et al. 2017c). In mmWave MIMO-NOMA, more than one users can be simultaneously supported in each beam via superposition coding and SIC operations. This is different from the conventional SDMA, where the signal for one user is only transmitted in each beam (Ding et al. 2017; Wang et al. 2017b). On the other hand, the small amount of RF chains in mmWave hybrid massive MIMO systems further limits the number of available beams and supported SDMA users. Consequently, much more users can be served within the same system resource block for any given RF chains via mmWave MIMO-NOMA,

Millimeter Wave Channel Access, Fig. 6

MIMO-NOMA system with two clusters: Each cluster includes two users covered by the same beam



and the system achievable sum-rate can be also improved Fig. 6.

The main challenges in mmWave MIMO-NOMA are given as follows (Xiao et al. [n.d.](#); Huang et al. [n.d.](#)). Similar to SISO-NOMA, SIC in mmWave MIMO-NOMA is implemented at the user side to suppress interference. While power allocation among users is the main focus in SISO-NOMA, it is possible to eliminate the user interference via beamforming and SIC in both the power and spatial domains in mmWave MIMO-NOMA, resulting in a more complicated design problem. For example, in SISO-NOMA, the SIC order generally depends directly on channel gains, e.g., a user with a smaller channel gain has a higher SIC order (i.e., decoded first). However, in mmWave MIMO-NOMA, effective channel gains are affected by beamforming designs. This results in a coupled beamforming and SIC ordering optimization. Moreover, some features in mmWave massive MIMO systems, e.g., massive antenna arrays, channels with limited scattering, and hybrid beamforming structures, make existing designs for the conventional MIMO-NOMA less applicable.

Due to the coarse hybrid beamforming capability and limited RF chains in mmWave systems, it is hard to serve many users only based on spatial resolution, especially for users with near directions. Introducing NOMA into mmWave MIMO can provide possibility of simultaneously serving these users. This advantage is of significant importance for those scenarios where both high data rate

and massive connectivity, e.g., dense urban scenarios, are highly desired. Furthermore, the increased path-loss component (Naqvi and Hassan [2016](#)) and channel blocking effect in mmWave communications are both beneficial for NOMA performance since they result in a larger channel gain difference for users covered by one beam with near locations. Considering the abovementioned two harmonies between the mmWave systems and the principle of NOMA, the use of NOMA for mmWave communications is a promising research direction.

Random Access in mmWave Systems

The basic procedure of random access in mmWave systems is similar to that of the traditional Sub-6 GHz cellular network, as shown in Fig. 3. However, in an mmWave cellular network, highly directional beamforming should be used at both the BS and users to compensate the large channel path loss. Since random access cannot benefit from the full beamforming gain due to the lack of information on the best transmit-receive beam pair (i.e., the location of strongest channel scatterer), the design of the random access becomes more challenging in mmWave communication systems (Jeong et al. [2015](#); Polese et al. [2017](#)). It has been shown that the exhaustive directional beam pair search is able to achieve a better preamble detection than the omnidirectional beam at the same cost of time overhead (Jeong et al. [2015](#)). Furthermore, reducing the time overhead of random access can be achieved by, e.g., (1) using enhanced preamble

detection method; (2) steering multiple receive beams in the same direction simultaneously to increase power of the received signals via noncoherently accumulation; (3) exploiting beam reciprocity to utilize the downlink beam information for the uplink; and (4) having better cell deployments for stronger channels (Jeong et al. 2015). Some other approaches directly aim to reduce the time overhead of the naive exhaustive search in random access for mmWave systems. For example, the angular space can be scanned in time-varying random directions using synchronization signals periodically transmitted by the BSs (Barati et al. 2015), the two-stage hierarchical procedure can be employed to speed up user discovery (Desai et al. 2014), and the context information on users and/or BS positions provided by a separate control plane can be also used to accelerate the random access procedure (Shokri-Ghadikolaei et al. 2015).

Key Applications

TDMA has been adopted in 802.11ad WLAN standard and 802.15.3c WPAN standard; spatial sharing multiple access has been used in 802.11ad WLAN standard; SDMA (multiple-user MIMO) has been added into 802.11aj WLAN standard (Complete Proposal for IEEE 802.11aj (45 GHz) n.d.; Hong et al. 2018) and is likely to be used in 802.11ay WLAN standard (an improved version of 802.11ad) (http://www.ieee802.org/11/Reports/tgay_update.htm).

Cross-References

- ▶ [Millimeter Wave Beam Training and Tracking](#)
- ▶ [Millimeter Wave Channel Modeling](#)
- ▶ [Millimeter Wave Massive MIMO](#)

References

- 802.11ad 2012 – IEEE Standard for Information technology – Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements–Part 11: Wireless

LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band. [Online] Available: <http://ieeexplore.ieee.org/document/6178212/>

- Alkhateeb A, Leus G, Heath RW (2015) Limited feedback hybrid precoding for multi-user millimeter wave systems. *IEEE Trans Wirel Commun* 14(11):6481–6494
- Barati CN et al (2015) Directional cell discovery in millimeter wave cellular networks. *IEEE Trans Wirel Commun* 14(12):6664–6678
- Baykas T et al (2011) IEEE 802.15. 3c: the first IEEE wireless standard for data rates over 1 Gb/s. *IEEE Commun Mag* 49(7):114–121
- Complete Proposal for IEEE 802.11aj (45 GHz). [Online] Available: https://mentor.ieee.org/802.11/documents?is_dcn=0707
- Dai L, Wang B, Yuan Y, Han S, Li C, Wang Z (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun Mag* 53(9):74–81
- Desai V, Krzymien L, Sartori P, Xiao W, Soong A, Alkhateeb A (2014) Initial beamforming for mmWave communications. In: *Proceedings of 48th Asilomar conference signals, system and computers*. IEEE, pp 1926–1930
- Ding Z, Yang Z, Fan P, Poor HV (2014) On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Signal Process Lett* 21(12):1501–1505
- Ding Z, Fan P, Poor HV (2017) Random beamforming in millimeter-wave NOMA networks. *IEEE Access* 5:7667–7681
- Goldsmith A (2005) *Wireless communications*. Cambridge University Press, New York
- Hany A, Widmer J (2017) Extending the IEEE 802.11 ad model: scheduled access, spatial reuse, clustering, and relaying In: *Proceedings of workshop on ns-3 (2017 WNS3)*, Porto, 13–14 June 2017, 8 pp. <https://doi.org/10.1145/3067665.3067667>
- He S, Wang J, Huang Y, Ottersten B, Hong W (2017a) Codebook based hybrid precoding for millimeter wave multiuser systems. *IEEE Trans Signal Process* 65(20):5289–5304
- He S, Wu Y et al (2017b) Joint optimization of analog beam and user scheduling for millimeter wave communications. *IEEE Commun Lett* 21(12):2638–2641
- Hong W, He S, Wang H, Yang G, Huang Y, Chen J, Yang L (2018) An overview of China millimeter-wave multiple gigabit wireless local area network system. *IEICE Trans Commun* 101(2):262–276
- Huang Y, Zhang C, Wang J, Jing Y, Yang L, You X Signal processing for MIMO-NOMA: present and future challenges. *IEEE Wireless Commun Mag*, 25(2): 32–38
- Jeong C, Park J, Yu H (2015) Random access in millimeter-wave beamforming cellular networks: issues and approaches. *IEEE Commun Mag* 53(1): 180–185

- Li Z, Han S, Molisch AF (2017) Optimizing channel-statistics-based analog beamforming for millimeter-wave multi-user massive MIMO downlink. *IEEE Trans Wirel Commun* 16(7):4288–4303
- Naqvi SAR, Hassan SA (2016) Combining NOMA and mmWave technology for cellular communication. In: *Proceedings of the IEEE vehicular technology conference (VTC Fall)*, Montreal. IEEE
- Pisinger D (2005) Where are the hard knapsack problems? *Comput Oper Res* 32:2271–2284
- Polese M, Giordani M, Mezzavilla M, Rangan S, Zorzi M (2017) Improved handover through dual connectivity in 5G mmWave mobile networks. *IEEE J Sel Areas Commun* 35(9):2069–2084
- Qiao J, Cai LX, Shen X, Mark JW (2012) STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks. In: *2012 IEEE International Conference on Communications (ICC)*, Ottawa. IEEE
- Roy RH III Ottersten B Spatial division multiple access wireless communication systems, U.S. Patent 5515378, 7 May 1991
- Rusek F et al (2013) Scaling up MIMO: opportunities and challenges with very large arrays. *IEEE Signal Process Mag* 30(1):40–60
- Shokri-Ghadikolaei H, Fischione C, Fodor G, Popovski P, Zorzi M (2015) Millimeter wave cellular networks: a MAC layer perspective. *IEEE Trans Commun* 63(10):3437–3458
- Sum C, Lan Z, Funada R, Wang J, Baykas T, Rahman MA, Harada H (2009) Virtual time-slot allocation scheme for throughput enhancement in a millimeter-wave multi-Gbps WPAN system. *IEEE J Sel Areas Commun* 27(8):1379–1389
- Sun S, Rappaport TS, Heath RW, Nix A, Rangan S (2016) MIMO for millimeter-wave wireless communications: beamforming, spatial multiplexing, or both? *IEEE Commun Mag* 52(12):110–121
- Technical Specifications and Technical Reports for a UTRAN-based 3GPP system, 3GPP TR 21.101. <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=544>
- Wang J, Peng Q, Huang Y, Wang H, You X (2017a) Convexity of weighted sum rate maximization in NOMA systems. *IEEE Signal Process Lett* 24(9):1323–1327
- Wang B, Dai L, Wang Z, Ge N, Zhou S (2017b) Spectrum and energy efficient beam-space MIMO-NOMA for millimeter-wave communications using lens antenna array. *IEEE J Sel Areas Commun* 35(10):2370–2382
- Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: enhancements for very high throughput for operation in bands below 6GHz, standard IEEE P802.11ac, Draft 0.1, IEEE Computer Society, Jan 2011
- Xiao Z, Dai L, Xia P, Choi J, Xia X Millimeter-wave communication with non-orthogonal multiple access for 5G, *IEEE Wireless Commun Mag*, to be published. [Online] Available: <https://arxiv.org/abs/1709.07980>
- Zhang C, Jing Y, Huang Y, Yang L (2017a) Performance scaling law for multi-cell multi-user massive MIMO. *IEEE Trans Veh Technol* 66(11):9890–9903
- Zhang C, Huang Y, Jing Y, Jin S, Yang L (2017b) Sum-rate analysis for massive MIMO downlink with joint statistical beamforming and user scheduling. *IEEE Trans Wirel Commun* 16:2181–2194
- Zhang D, Zhou Z, Xu C, Zhang Y, Rodriguez J, Sato T (2017c) Capacity analysis of non-orthogonal multiple access with mmwave massive MIMO systems. *IEEE J Sel Areas Commun* 35(7):1606–1618
- Zhang C, Jing Y, Huang Y, Yang L (2018) Interleaved training and training-based transmission design for hybrid massive antenna downlink. *IEEE J Select Topics Signal Process* <https://doi.org/10.1109/JSTSP.2018.2818648>
- Zhu J, Wang J, Huang Y, He S, You X, Yang L (2017) On optimal power allocation for downlink non-orthogonal multiple access systems. *IEEE J Sel Areas Commun* 35(12):2744–2757

Millimeter Wave Channel Measure

Sherif Adeshina Busari¹, Shahid Mumtaz¹, Kazi Mohammed Saidul Huq¹, and Jonathan Rodriguez^{1,2}

¹Instituto de Telecomunicações, Aveiro, Portugal

²University of South Wales, Pontypridd, UK

Synonyms

mmWave channel campaigns; mmWave channel measurement; mmWave channel sounding

Definitions

Channel measurement refers to the techniques widely used in telecommunications for studying the properties of wireless channels. It is done by carrying out real-world measurements using a set of specialized equipment known as channel sounders (transmitter(s) and receiver(s)). Typically, the field measurements are undertaken at different times, cities, environments, and under

different scenarios, resulting in what is referred to as channel measurement campaigns. The results from the measurements are used to develop channel models for the characterization and performance evaluation of wireless systems and networks. Millimeter wave (mmWave) channel measurements thus refer to measurements carried out at mmWave frequencies (30–300 GHz) (Busari et al. 2017).

Key Points

The mmWave frequency bands fall in the extremely high frequency (EHF) range of the electromagnetic spectrum. The bands have frequencies (f) spanning [30, 300] GHz, which corresponds to wavelengths (λ) of [10, 1] mm, respectively, as $f = c/\lambda$ and $c = 3 \times 10^8$ m/s is the speed of light (Busari et al. 2017). While this is the general range for mmWave, some authors in the literature limit the mmWave frequencies to the 30–90/100 GHz bands, preferring to refer to the 90/100–300 GHz frequencies as part of the terahertz (THz) bands (Mumtaz et al. 2017a). In addition, frequencies above 6 GHz but lower than 30 GHz, such as frequencies between 10 and 28 GHz, are also referred to as mmWave (Xiao et al. 2017). In each case, the bands have wavelengths up to a few millimeters, hence the name mmWave.

Relative to the sub-6 GHz microwave (μ Wave) frequencies used for many common applications, the mmWave bands offer several advantages. The larger bandwidths translate to higher capacity, the smaller wavelengths enable more (i.e., massive) antennas in same physical space, and the relatively closer spectral allocations lead to a more homogeneous propagation (Busari et al. 2017).

On the other hand, mmWave signals are prone to higher path loss, higher penetration loss, severe atmospheric absorption, higher susceptibility to shadowing and diffraction, more attenuation due to rain, and higher vulnerability to blockages by objects, as their wavelengths are typically less than the physical dimensions of the obstacles (Busari et al. 2017; Rappaport et al. 2013).

To counter the effects of the propagation losses, reduce interference, increase system throughput, and maximize network's energy efficiency, directional communication is mandatory for practical mmWave systems. Thus, beamforming can be employed to increase the antenna array gains, suppress interference, and improve the signal to noise ratio (SNR) of the links (Huang et al. 2018; Cui et al. 2018).

In addition, relay stations can be used to circumvent obstacles, thereby avoiding blockages and minimizing outages (Yang et al. 2018). More so, for the 50–200 m cell size envisaged for mmWave small cells (and other mmWave short-range applications), the expected attenuation (about 7 dB/km), due to heavy rainfalls, has minimal effect within such distances. Also, the high mmWave path loss limits inter-cell interference (ICI) and allows for more frequency reuse, which by extension improves the overall system capacity (Rappaport et al. 2013).

Historical Background

The study of mmWave predates the twentieth century. In the 1890s, scientists performed experiments at mmWave frequencies. However, mmWave channel measurements and system designs targeted at applications for radio communication started in the 1990s and early 2000s (Xiao et al. 2017). The spectrum crunch at μ Wave and the projected explosive traffic demands with the coming of fifth-generation (5G) networks and the Internet of Things (IoT)/Internet of Everything (IoE) have again motivated rigorous research and development (R&D) for mmWave systems and networks (Busari et al. 2017).

Foundations

MmWave channel measurement has received tremendous attention in recent years. Researchers in the academia and the industry have carried

out extensive mmWave measurement campaigns in different cities, countries, and continents of the world. A summary of mmWave channel measurements in the major cities of Europe, Asia, and the United States of America between 2012 and 2017 can be found in Xiao et al. (2017).

Since 2011, the New York University (NYU) WIRELESS researchers have been conducting extensive measurements at 28, 38, 60, and 73 GHz (Rappaport et al. 2013, 2015). The EU mmMAGIC project team also carried out mmWave measurements for mmWave frequencies 10–100 GHz (Jaeckel et al. 2014). Others include Zhao et al. at 32 GHz Zhao et al. (2017), Ai et al. at 26 GHz [8], METIS at 50–70 GHz (Ai et al. 2017), and the 5GCM at 6–100 GHz (Ghosh 2015), among others.

While most existing mmWave channel measurements were done for static radio channels, the world's first demonstration for the dynamic channel was carried out by Samsung at 28 GHz in 2014, reaching 1.2 Gbps transmission rate at 110 kmph. Contrary to earlier thoughts, this dynamic channel measurement points to the suitability of mmWave frequencies for mobile applications. It is also particularly useful to buttress the viability of mmWave small cells which typically do not involve high-velocity users (Mumtaz et al. 2017b).

For good mmWave channel measurements, the sounding equipment (both hardware and software) must have high performance and reliability. This is essential for the accuracy of the measurements and the corresponding models derived from them. High-frequency (mmWave) bands place high demands on the sounder components, as they must ensure wide frequency coverage, high dynamic range, and high output signal (in terms of power, stability, and quality) as well as guarantee the least possible distortion and harmonic content (Busari et al. 2017; Mumtaz et al. 2017b).

Efficient channel sounders must therefore be capable of measurements with operating frequencies up to several tens/hundreds of GHz and several GHz of mmWave bandwidth as well as multiple antennas at the transceivers. Similarly, sounding equipment should be capable of

measurements in static and dynamic conditions and indoor and outdoor environments. They should as well have capabilities for efficient signal generation, data acquisition, and storage. The equipment should also guarantee accurate synchronization, calibration, sensitivity, resolution, and flexibility for extension while maintaining reasonable system cost (Busari et al. 2017; Mumtaz et al. 2017b).

The extensive mmWave channel measurements have revealed marked differences in the propagation characteristics of mmWave relative to μ Wave. MmWave signals exhibit line-of-sight (LOS) or near-LOS propagation with absolute increase in path loss with increasing frequency, specular reflection attenuation, diffuse scattering, very high diffraction attenuation, and frequency dispersion effect, which, with the prospect of the huge bandwidth, allows the propagation to be considered as frequency-dependent. The mmWave signals also exhibit quasi-optical properties as the short wavelengths lead to weak diffractions. With highly directional mmWave communications, delay spreads of less than 20 ns are typical, and the number of scattering clusters becomes much less than at μ Wave (Rappaport et al. 2015; Xiao et al. 2017; Busari et al. 2017).

Under static conditions, mmWave channels exhibit frequency-selective fading which would need to be addressed by modulation or equalization. At mmWave frequencies, the assumption of asymptotic pair-wise orthogonality between channel vectors under independent and identically distributed (i.i.d) Rayleigh fading channel which is valid for μ Wave bands no longer holds. The number of independent multipath components (MPCs) becomes limited, and so channel vectors exhibit correlated fading. The channels are non-orthogonal and the waves are spherical. Spatial non-stationarity also becomes severe. Extensive indoor and outdoor channel measurement campaigns and simulations have been carried out by Rappaport et al. (2013), Rappaport et al. (2015), Jaeckel et al. (2014), Zhao et al. (2017), Ai et al. (2017), METIS (2015), and Ghosh (2015), among others, to deduce these outcomes.

Following the mmWave channel measurement activities, channel parameters (such as path loss, penetration loss, power delay profile (PDP), delay spread, shadowing, angular spreads, coherence bandwidth, etc.) are estimated from the obtained data or field results to develop channel models. Considering all necessary factors (such as frequency, propagation environment, scenario, etc.), the developed channel models provide statistical, mathematical, and analytical frameworks for simulation studies and performance evaluation of the communication systems. Such frameworks are also used to compare and/or validate empirical data from field deployment and operation tests (Mumtaz et al. 2017b).

Some of the general mmWave channel models that have resulted from mmWave measurement efforts include MiWEBA, COST 2100, QuaDRiGa, mmMAGIC, METIS, NYUSIM, 3GPP TR 38.900, 5GCM SIG, IMT-2020, and MG5GCM channel models, among others (3GPP 2016).

Key Applications

Due to the attractive capabilities and high commercial potentials of mmWave communications, spectrum allocation, regulatory, and standardization activities are being undertaken and rigorously pursued by relevant bodies such as the ITU, IEEE, ETSI, and the 3GPP. In fact, the 60 GHz mmWave WiFi (IEEE 802.11ad) that supports transmission rates up to 4–7 Gbps has already been standardized and has hit the market. It is expected to be followed by the 60 GHz mmWave IEEE 802.11ay capable of higher data rates up to 20–40 Gbps (Xiao et al. 2017).

5G mmWave cellular networks are also anticipated to be deployed by year 2020 to support many high data rate applications such as short-range communications, vehicular networks, and wireless in-band fronthauling/backhauling, among others. As a result, measurement campaigns, channel modeling, demos, prototyping, and field tests for mmWave technologies and applications have intensified. Standardization,

holistic deployment, performance evaluation, and business models will follow these efforts (Busari et al. 2017).

Future Directions

The mmWave channel is intrinsically an ultra-broadband channel with huge bandwidth and high spatial multiplexing capabilities to significantly enhance wireless access and improve cell and user throughputs. The mmWave spectrum has abundant available bandwidth (up to 270 GHz). With a reasonable assumption of 40% availability over time, the bands will possibly open up about 100 GHz new spectrum for different applications and services (such as cellular and fixed wireless communications) (Busari et al. 2017; Liu et al. 2018).

Since there is spectrum shortage at the congested sub-6 GHz μ Wave bands, relative to the explosive data rate demands projected for 2020 and beyond, mmWave wireless offers promising and exciting potentials to support next-generation systems. As a result, it features in the list of key enablers for the multi-gigabit-per-second (Gbps) wireless access for the fifth-generation (5G) and beyond-5G (B5G) mobile networks, as well as for high-rate wireless personal and local area networks (WPAN and WLAN) (Yang et al. 2018).

Though there are many challenges to address, the mmWave technology shows amazing prospects and potentials that will undoubtedly usher in a new era for services and applications not hitherto possible. Certainly, mmWave channel measurement is a critical component of the exciting journey.

Cross-References

- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)
- [Millimeter-wave Communications](#)

References

- 3GPP (2016) Study on channel model for frequency spectrum above 6 GHz (3GPP TR38.900), v14.2.0. Technical report, 3GPP. http://www.3gpp.org/ftp/Specs/archive/38_series/38.900/
- Ai B, Guan K, He R, Li J, Li G, He D, Zhong Z, Huq KMS (2017) On indoor millimeter wave massive MIMO channels: measurement and simulation. *IEEE J Sel Areas Commun* 35(7):1678–1690. <https://doi.org/10.1109/JSAC.2017.2698780>
- Busari SA, Huq KMS, Mumtaz S, Dai L, Rodriguez J (2017) Millimeter-Wave Massive MIMO Communication for Future Wireless Systems: A Survey. *IEEE Communications Surveys & Tutorials*, 20(2):836–869, Secondquarter 2018. <https://doi.org/10.1109/COMST.2017.2787460>
- Cui Y, Fang X, Fang Y, Xiao M (2018) Optimal Nonuniform Steady mmWave Beamforming for High-Speed Railway. *IEEE Transactions on Vehicular Technology*, 67(5):4350–4358. <https://doi.org/10.1109/TVT.2018.2796621>
- Ghosh A (2015) 5G channel model for bands up to 100 GHz. In: *IEEE Globecom 2015*. IEEE Globecom, San Diego. <http://www.5gworkshops.com/5GCM.html>
- Huang Y, Zhang J, Xiao M (2018) Constant Envelope Hybrid Precoding for Directional Millimeter-Wave Communications. *IEEE Journal on Selected Areas in Communications*. <https://doi.org/10.1109/JSAC.2018.2825820>
- Jaekel S, Raschowski L, Borner K, Thiele L (2014) QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials. *IEEE Trans Antennas Propag* 62(6):3242–3256. <https://doi.org/10.1109/TAP.2014.2310220>
- Liu Y, Fang X, Xiao M, Mumtaz S (2018) Decentralized beam pair selection in multi-beam millimeter-wave networks. *IEEE Trans Commun* PP(99):1–1. <https://doi.org/10.1109/TCOMM.2018.2800756>
- METIS (2015) Channel models deliverable 1.4 version 3 document ICT-317669 METIS/D14. Technical report, METIS
- Mumtaz S, Jornet JM, Aulin J, Gerstacker WH, Dong X, Ai B (2017a) Terahertz communication for vehicular networks. *IEEE Trans Veh Technol* 66(7):5617–5625
- Mumtaz S, Rodriguez J, Dai L (eds) (2017b) *mmWave massive MIMO: a paradigm for 5G*, 1st edn. Academic Press, London/San Diego
- Rappaport TS, Sun S, Mayzus R, Zhao H, Azar Y, Wang K, Wong GN, Schulz JK, Samimi M, Gutierrez F (2013) Millimeter wave mobile communications for 5G cellular: it will work! *IEEE Access* 1:335–349. <https://doi.org/10.1109/ACCESS.2013.2260813>
- Rappaport TS, MacCartney GR, Samimi MK, Sun S (2015) Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design. *IEEE Trans Commun* 63(9):3029–3056. <https://doi.org/10.1109/TCOMM.2015.2434384>
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis GK, Bjornson E, Yang K, Chih-Lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun* 35(9):1909–1935. <https://doi.org/10.1109/JSAC.2017.2719924>
- Yang G, Xiao M, Al-Zubaidy H, Huang Y, Gross J (2018) Analysis of millimeter-wave multi-hop networks with full-duplex buffered relays. *IEEE/ACM Trans Netw* 26(1):576–590. <https://doi.org/10.1109/TNET.2017.2786341>
- Yang G, Xiao M, Alam M, Huang Y (2018) Low-latency heterogeneous networks with millimeter-wave communications. *CoRR*. arXiv preprint arXiv:180109286
- Zhao X, Li S, Wang Q, Wang M, Sun S, Hong W (2017) Channel measurements, modeling, simulation and validation at 32 GHz in outdoor microcells for 5G radio systems. *IEEE Access* 5:1062–1072. <https://doi.org/10.1109/ACCESS.2017.2650261>

Millimeter Wave Channel Modeling

Wei Huang¹, Jie Huang², Yongming Huang³, and Cheng-Xiang Wang¹

¹Southeast University, Nanjing, China

²Shandong University, Qingdao, China

³School of Information Science and Engineering, Southeast University, Nanjing, China

Synonyms

[Millimeter wave channel modeling](#)

Definition

Millimeter wave channel modeling is to use various deterministic or stochastic channel modeling approaches to describe the channel propagation characteristics at millimeter wave frequency bands and verify them by channel measurements.

Historical Background

Channel modeling is important for wireless communication system design, test, and performance

evaluation. It has been the foundation of wireless communications for years. As wireless communication systems evolve from the first generation (1G) to the forthcoming fifth generation (5G), many channel models have been proposed, among which the series of COST-, WINNER-, and 3GPP-family models are the most widely investigated ones. The description of wireless channels has also evolved from time domain to time, delay, and spatial domains. Another obvious distinction between different generations of wireless communication systems lies in the carrier frequency.

Due to the congestion of sub-6 GHz bands, millimeter wave (mmWave) technology has been proposed as a key technology for 5G to enable high data rate transmissions (Wang et al. 2014). In general, it denotes the frequency range of 30–300 GHz, but sometimes 10–30 GHz bands are also included as they share some similar propagation characteristics (Huang et al. 2017). MmWave channel measurements and modeling have drawn much attention since T. S. Rappaport's breakthrough work in Rappaport (1991). It demonstrated that mmWave can work for 5G mobile cellular networks.

In the early stage, research of mmWave concentrated on 60 GHz bands as there are at least 5 GHz bandwidths worldwide (Wu et al. 2017a). Recently, other frequency bands such as 11, 15, 26, 28, 32, 38, and 73 GHz have widely been investigated (Rappaport et al. 2015). Various channel measurements and models at these frequency bands have pushed the application of mmWave communications to 5G wireless networks.

Moreover, channel models are required for simulating propagation in a reproducible and cost-effective way and are used to accurately design and compare radio air interfaces and system deployments. Common wireless channel model parameters include carrier frequency, bandwidth, two-dimensional (2-D) or three-dimensional

(3-D) distance between the transmitter (Tx) and receiver (Rx), Doppler effects, delays, and angles. The definitive challenge for a 5G channel model is to provide a fundamental physical basis while

being flexible and accurate, especially across a wide frequency range such as 0.5–100 GHz. Recently, a lot of research works aiming at understanding the propagation mechanisms and channel behavior at the frequency bands above 6GHz have been published.

Foundations

For mmWave channel modeling, some fundamental knowledge will be useful. MmWave frequency allocations, propagation characterizations, channel model standardizations, channel modeling methods, and proposed mmWave channel models will be discussed here.

Frequency Allocations

The 57–64 GHz band was allocated in North America and Korea. In Japan, Europe, China, and Australia, the 59–66, 57–66, 59–64, and 59.4–62.9 GHz bands were allocated, respectively. At the World Radio Communication Conference 2015 (WRC-15), 11 frequency bands between 24.25 GHz and 86 GHz are allocated for 5G by ITU. The 24.25–27.5, 37–40.5, 42.5–43.5, 45.5–47, 47.2–50.2, 50.4–52.6, 66–76, and 81–86 GHz bands were allocated, and the 31.8–33.4, 40.5–42.5, and 47–47.2 GHz bands require additional allocations. Those frequency allocations have accelerated the developments of mmWave communications around the world.

In the United States, the 27.5–28.35 (28), 37–38.6 (37), 38.6–40 (39), and 64–71 GHz bands were allocated for flexible wireless use. Additional allocated bands include 24.25–24.45, 24.75–25.25 (24), and 47.2–48.2 GHz. Europe limits its considerations to the bands listed by WRC-15, in particular the 24.25–27.5 (26), 31.8–33.4 (32), and 40.5–43.5 GHz bands. In Korea, the 26.5–29.5 GHz (28 GHz) band was allocated. In Japan, the 27.5–29.5GHz (28 GHz) band was allocated in accordance with America and Korea. In China, the focused bands are 24.25–27.5 GHz (26GHz) and 37–43.5 GHz bands.

Propagation Characterizations

MmWave has some unique propagation characteristics, which should be carefully studied (Hemadep et al. 2018). MmWave has high path loss and high penetration loss, especially for outdoor-to-indoor (O2I) scenarios. Atmospheric attenuation becomes severe at mmWave bands. The attenuation caused by oxygen absorption can be 15 dB/km and 1.4 dB/km for 60 GHz and 120 GHz, respectively. For water vapor, the attenuation would be 0.18 dB/km and 28.35 dB/km at 23 GHz and 183 GHz, respectively. Rain can also cause 2.55–20 dB/km attenuation depending on the rainfall rate. Foliage also adds additional attenuation. As the size of many objects is at the level of the wavelength, diffraction becomes an important propagation mechanism, and the first Fresnel zone is narrow. The channel also shows sparsity in spatial domain. Blockage caused by human movement in indoor scenarios and buildings in outdoor scenarios will make the line-of-sight (LOS) component unfeasible, thus changing the channel from LOS to non-LOS (NLOS) scenarios (Qi et al. 2017).

Channel Model Standardizations

Several standards have dedicated to 60 GHz bands. IEEE 802.15.3c and 802.11ad have been developed for wireless personal area network (WPAN) in March 2007 and wireless local area network (WLAN) in May 2010, respectively, while 802.11ay is being developed for next-generation wireless fidelity (WiFi). MiWEBA channel model was released in June 2014 for 60 GHz (57–66 GHz) outdoor scenarios, including open area (campus), street canyon, hotel lobby, fronthaul/backhaul, and device-to-device (D2D).

The METIS channel model was released in February 2015. It can support up to 86 GHz with high bandwidth. Supported scenarios include dense urban, urban, rural, indoor (office and shopping mall), and highway.

The mmMAGIC channel model was released in May 2017. Its frequency range is from 6 GHz to 100 GHz. Supported scenarios include urban micro (UMi) (street canyon and open square),

indoor (office, shopping mall, and airport), O2I, stadium, and metro station.

The 5GCM white paper was first released in December 2015 in IEEE Globecom'15 and updated in October 2016. It intends to develop channel models for frequency bands up to 100 GHz using the geometry-based stochastic channel modeling (GBSM) approach and has conducted many channel measurements in various scenarios including UMi, urban macro (UMa), and indoor. The majority of the modeling ideas and parameter values were adopted in 3GPP TR 38.900 named “Study on Channel Model for Frequency Spectrum Above 6 GHz”, which was first released in February 2016 and updated in July 2017. Its alternative version is 3GPP TR 38.901 named “Study on Channel Model for Frequencies From 0.5 to 100 GHz”, in which the sub-6 GHz bands are also covered. It was first released in February 2017 and updated in December 2017.

Channel Modeling Methods

In general, channel modeling methods can be classified into deterministic and stochastic ones or the combination of them (Yang et al. 2017). Deterministic channel models are site-specific and can be verified by comparison of detailed path parameters with channel measurements. Stochastic channel models aim to model general environments and can be verified by comparison of channel statistical properties with channel measurements.

Deterministic channel models include ray tracing model, map-based model, and point cloud model. Ray tracing is based on geometry optics (GO), geometrical theory of diffraction (GTD), and uniform theory of diffraction (UTD), which are approximation and simplification of high-frequency electromagnetic propagations mechanisms. The METIS map-based model is based on ray tracing using a simplified geometric description of the environment. Point cloud is a ray tracing-like prediction tool to characterize the environment with higher precision (Järveläinen et al. 2016). Quasi-deterministic (Q-D) model, which is adopted by MiWEBA and IEEE 802.11ay, is a semi-deterministic channel model.

Stochastic channel models include SV-based model, propagation graph model, and GBSM. SV-based model assumes rays arrive in clusters and is used in IEEE 802.15.3c and 802.11ad models. Propagation graph model is based on graph theory. It models Tx, Rx, and scatterers as vertices and propagation paths as edges. GBSM has widely been used for channel modeling and adopted in many standard channel models like WINNER, 3GPP TR 38.900/901, NYU WIRELESS, mmMAGIC QuaDRiGa, and METIS stochastic models.

Proposed mmWave Channel Models

In order to meet the requirements of mmWave channel models, many major organizations, such as 3GPP, METIS, mmMAGIC, and 5GCM, have proposed the models for LOS probability, path loss, and building penetration. In the LOS probability model, the LOS probability is often modeled as a function of the 2-D distance between the Tx and Rx, and it includes UMi LOS probability, UMa LOS probability, and InH LoS probability. Furthermore, the large-scale path loss and O2I penetration loss models were also proposed in Rappaport et al. (2017).

Key Applications

As we know, 5G will be used in three scenarios, i.e., enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultrareliable low-latency communications (uRLLC). MmWave will be a key technology to enable these applications. Typical and potential application scenarios include fronthaul/backhaul, indoor hotspots, cellular networks, high-speed train (HST), vehicle-to-vehicle (V2V), D2D, unmanned aerial vehicle (UAV) communications, etc.

Future Directions

mmWave Channel Measurements in New Scenarios

As mmWave will be applied to various new scenarios, channel measurement campaigns in

these challenging scenarios are indispensable, which include but are not limited to massive multiple-input multiple-output (MIMO) (Busari et al. 2017), HST, V2V, D2D, UAV communications, etc. MmWave channel measurements in these new scenarios will lead to a better understanding of mmWave propagation characteristics and more accurate mmWave channel models.

General 5G Channel Models

Though many channel models have been proposed, most of them cannot fulfill all the requirements of 5G, covering the vast mmWave frequency bands and new propagation characteristics. A preliminary general 5G channel model was proposed in Wu et al. (2017b). It aimed at capturing small-scale fading channel characteristics of key 5G communication scenarios, such as massive MIMO, HST, V2V, and mmWave. It will serve as a good guideline for future more general and standard 5G models.

Big Data-Enabled mmWave Channel Modeling

Wireless big data has now been investigated both in academia and industry to handle massive datasets generated in wireless networks using artificial intelligence (AI) technologies and machine learning (ML) algorithms (Bi et al. 2015). Big data analytics have been applied to different layers of wireless networks for channel estimation, positioning/localization, network optimization and management, etc. Some preliminary works also apply big data analytics to wireless channel modeling (Ma et al. 2017). Big data will be a powerful tool for mmWave channel modeling by training and learning big channel measurement or simulation datasets.

References

- Bi S, Zhang R, Ding Z, Cui S (2015) Wireless communications in the era of big data. *IEEE Commun Mag* 53(10):190–199
- Busari SA, Huq KMS, Mumtaz S, Dai L, Rodriguez J (2017) Millimeter-wave massive MIMO communi-

- cation for future wireless systems: a survey. *IEEE Commun Surv Tutorials* PP(99):1–1
- Hemaddeh I, Satyanarayana K, El-Hajjar M, Hanzo L (2018) Millimeter-wave communications: physical channel models, design considerations, antenna constructions and link-budget. *IEEE Commun Surv Tutorials* PP(99):1–1
- Huang J, Wang CX, Feng R, Sun J, Zhang W, Yang Y (2017) Multi-frequency mmWave massive MIMO channel measurements and characterization for 5G wireless communication systems. *IEEE J Sel Areas Commun* 35(7):1591–1605
- Järveläinen J, Haneda K, Karttunen A (2016) Indoor propagation channel simulations at 60 GHz using point cloud data. *IEEE Trans Antennas Propag* 64(10):4457–4467
- Ma X, Zhang J, Zhang Y, Ma Z, Zhang Y (2017) A PCA-based modeling method for wireless MIMO channel. In: 2017 IEEE conference on computer communications workshops (INFOCOMWKSHPS), Atlanta, pp 874–879
- Qi W, Huang J, Sun J, Tan Y, Wang CX, Ge X (2017) Measurements and modeling of human blockage effects for multiple millimeter wave bands. In: 2017 13th international wireless communications and mobile computing conference (IWCMC), Valencia, pp 1604–1609
- Rappaport TS (1991) The wireless revolution. *IEEE Commun Mag* 29(11):52–71
- Rappaport TS et al (2015) Millimeter wave wireless communications. Prentice-Hall, upper saddle river
- Rappaport TS et al (2017) Overview of millimeter wave communications for fifth-generation (5G) wireless networks with a focus on propagation models. *IEEE Trans Antennas Propag* 65(12):6213–6230
- Wang CX, Haider F, Gao X, You XH, Yang Y, Yuan D, Aggoune HM, Haas H, Fletcher S, Hepsaydir E (2014) Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Commun Mag* 52:122–130
- Wu X, Wang CX, Sun J, Huang J, Feng R, Yang Y, Ge X (2017a) 60-GHz millimeter-wave channel measurements and modeling for indoor office environments. *IEEE Trans Antennas Propag* 65(4):1912–1924
- Wu S, Wang CX, Aggoune EHM, Alwakeel MM, You XH (2017b) A general 3D non-stationary 5G wireless channel model. *IEEE Trans Commun* PP(99):1–1
- Yang Y, Xu J, Shi G, Wang CX (2017) 5G wireless systems: simulation and evaluation techniques, Wireless networks. Springer, Cham. URL <http://cds.cern.ch/record/2300921>

Millimeter Wave Communication Security

► Millimeter Wave Security

Millimeter Wave Large-Scale MIMO

► Millimeter Wave Massive MIMO

Millimeter Wave MAC Layer

Hossein S. Ghadikolaei¹ and Carlo Fischione²

¹Department of Networks and Systems Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

²KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Medium access control in millimeter-wave wireless communications

Definition

The millimeter-wave (mmWave) medium access control (MAC) is a set of routines that regulate access to a common wireless communications medium in the mmWave frequency bands. MAC layer is part of layer 2 of the OSI model (Bertsekas and Gallager 1992). The primary functions of this layer for mmWave wireless communications are the following:

- Data encapsulation, including the frame assembly and error control;
- Media access control, including initial frame transmission, retransmission in the case of failure, and flow control; and
- Initial access, mobility management, and handover.

The first two items are common to MAC of any communication systems, and the last one is more specific for mmWave networks.

The choice of MAC protocol directly influences most performance measures, including

end-to-end latency, network throughput, reliability, and energy efficiency, among others.

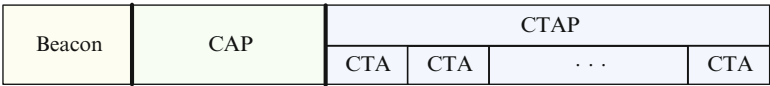
Historical Background

Millimeter-wave communications use the part of the electromagnetic spectrum in the range 30–300 GHz, which corresponds to wavelengths from 10 to 1 mm. In the literature, however, mmWave frequencies casually refer to the frequency band between 6 and 300 GHz. The mmWave wireless communications exhibit the following fundamental physical characteristics: short wavelengths, large bandwidth, severe propagation loss, high attenuation through most solid materials, sparse scattering environments, and large-scale antennas (Rappaport et al. 2013). The special characteristics of mmWave systems demand a thorough reconsideration of traditional protocol design principles, especially at the MAC layer with respect to the MAC protocols of traditional wireless communications.

When sending data over a network, the MAC layer encapsulates its incoming payload (data frames from higher layers) into frames appropriate for the transmission medium, adds a frame check sequence to identify transmission errors, and then forwards the data to the physical layer after running a multiple access procedure (Bertsekas and Gallager 1992). This procedure is necessary to resolve or avoid collisions, retransmit the collided or lost frames, and control the transmission rate. At the receiver side, the MAC layer verifies the sender's frame check sequences, removes preambles and frame headers, and delivers the payload to the higher layers (Bertsekas and Gallager 1992). In wireless networks, the design of the MAC layer is tightly integrated with the design of the physical layer. The multiple access protocol of short-range wireless networks (such as wireless personal and local area networks) may rely on carrier sensing among devices and multihop communications. The multiple access protocols of the cellular networks, like Long-Term Evolution (LTE), are usually optimized and run by the base stations (3GPP TS 36.321 2017).

At the mmWave frequency band, there are several standardization activities within cellular networks, wireless personal area networks (WPANs), and wireless local area networks (WLANs), such as 3GPP NR 3GPP TS 38.321 (2017), IEEE 802.15.3 Task Group 3c (TG3c) (IEEE 802.15.3c 2009), IEEE 802.11ad standardization task group (IEEE 802.11ad 2012), WirelessHD Consortium, and Wireless Gigabit Alliance (WiGig). The existing established standards are very suboptimal, and efficient MAC layer design for mmWave networks is a topic of intense research. Ghadikolaei et al. (2015) and Ghadikolaei et al. (2016) discussed the main MAC layer issues of infrastructure-based (e.g., cellular) and infrastructure-free (e.g., ad hoc) mmWave networks, respectively. For short-range networks, IEEE 802.11ay is the most recent study group within IEEE, formed in May 2015, which aims to modify IEEE 802.11ad to enhance the throughput, range, and most importantly the use cases, while ensuring backward compatibility and coexistence with legacy mmWave standards. Broadly speaking, IEEE standards define a network with one coordinator and several mmWave devices. The coordinator, which can be a device itself, is responsible for broadcasting, synchronization beacons, running initial access beacons, and managing radio resources. Figure 1 illustrates generic timing segmentation of IEEE 802.15.3c and IEEE 802.11ad.

The IEEE 802.15.3c MAC procedure was standardized in 2009 (IEEE 802.15.3c 2009). It selects one device, among a group of devices in the network, as the piconet coordinator (PNC), broadcasting beacon messages. Time is divided into successive super-frames, each consisting of three portions, beacon, contention access period (CAP), and channel time allocation period (CTAP), as shown in Fig. 1a. In the beacon, the coordinator transmits omnidirectional or multiple quasi-omnidirectional beacons to facilitate the device discovery procedure. In the CAP, devices with low QoS requirements start their data transmissions. Devices with high QoS requirements contend to register their channel access requests at the PNC, based on carrier



(a) Superframe of IEEE 802.15.3c



(b) Beacon interval of IEEE 802.11ad

Millimeter Wave MAC Layer, Fig. 1 Network timing structure of existing IEEE mmWave standards. In IEEE 802.15.3c, beacon messages are transmitted in the Beacon phase. Channel access requests are made in CAP and served in CTAP using TDMA. Similar procedures are adopted in IEEE 802.11ad. (This picture is reproduced from Ghadikolaei et al. 2016). (a) Superframe of IEEE 802.15.3c. (b) Beacon interval of IEEE 802.11ad

Millimeter Wave MAC Layer, Table 1 Application scenarios for short-/medium-range mmWave networks. This table is deduced from ongoing discussions inside IEEE 802.11ay study group. “NS” means not specified yet

Usage models	Delay (s)	Availability	Range (m)	Rate (Gbps)	Application scenarios
Ultrashort-range communications	<1	NS	<10	10	Wireless tollgate to transfer e-magazine, picture library, 4K movie trailers, 4K movies
8K Video transfer at smart home	<0.005	NS	<5	28	8K video stream between a source device (e.g., setup box) and a sink device (e.g. smart TV), replacement of wired interface
Augmented reality	<0.005	NS	<10	20	Interface between a mobile wearable device and its managing device to deliver 3D video
Data center	<0.1	99.99%	<5	40	Inter-rack connectivity
Vehicular networks	<0.1	NS	<1000	ns	Intra- and inter-car connectivity, intersection collision avoidance, public safety
Video on-demand	<0.1	NS	<100	NS	Broadcast in crowd public places (e.g., classroom, in flight, train, bus, exhibitions)
Mobile offloading	<0.1	99.99%	<100	20	Offload video traffic from cellular interface to the mmWave interface
Mobile fronthauling	<0.035	99.99%	<200	20	Wireless connections between remote radio heads and base band unit
Mobile backhauling	<0.035	99.99%	<1000	20	Small cell backhauling, multihop backhauling, inter-building communications

sense multiple access with collision avoidance (CSMA/CA). The PNC serves the registered devices during CTAP. Resource allocation in CTAP is based on time division multiple access (TDMA), in which dedicated channel time allocations (CTAs) are used for different data traffic.

The IEEE 802.11ad was standardized in 2012. It adds modifications to the IEEE 802.11 physical and MAC layers to enable mmWave communications at 60 GHz (IEEE 802.11ad 2012). It defines a network as a personal basic ser-

vice set (PBSS) with one coordinator, called PBSS control point (PCP), and several stations. A super-frame, called beacon interval, is divided into a beacon header interval (BHI) and a data transfer interval (DTI); see Fig. 1b. BHI consists of a beacon transmission interval (BTI), an association beamforming training (A-BFT), and an announcement transmission interval (ATI). DTI consists of several contention-based access periods (CBAPs) and service periods (SPs). In BTI, PCP transmits directional beacon frames that contain basic timing for the personal BSS, fol-

lowed by beamforming training and association to PCP in the A-BFT period. ATI is allocated for request-response services where PCP sends information to the stations. Depending on the required QoS level, a device will be scheduled in the CBAP to transmit data using CSMA/CA or in the SP for contention-free access using TDMA. This schedule is announced to the participating stations prior to the start of DTI.

Key Applications

Currently, the mmWave spectrum is primarily used for satellite communications, long-range point-to-point communications, military applications, and local multipoint distribution service (Rappaport et al. 2014). There are growing interests in using mmWave systems to support extremely high data rate and low latency services for short- and medium-range networks. Table 1 shows some of these potential use cases.

Cross-References

- ▶ [Interference Management](#)
- ▶ [MAC in Cognitive Radio Networks](#)
- ▶ [Millimeter-wave Communications](#)
- ▶ [Multiple Access Technique for Cellular Wireless Networks](#)
- ▶ [QoS-Aware MAC](#)
- ▶ [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

References

- 3GPP TS 36321 (2017) Evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification
- 3GPP TS 38321 (2017) NR; medium access control (MAC) protocol specification evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification
- Bertsekas DP, Gallager RG (1992) Data networks, vol 2. Prentice-Hall, Englewood Cliffs
- Ghadikolaei HS, Fischione C, Fodor G, Popovski P, Zorzi M (2015) Millimeter wave cellular networks: a MAC layer perspective. *IEEE Trans Commun* 63(10): 3437–3458
- Ghadikolaei HS, Fischione C, Popovski P, Zorzi M (2016) Design aspects of short range millimeter wave wireless networks: a MAC layer perspective. *IEEE Netw* 30(3):88–96
- IEEE 802.11ad (2012) IEEE 802.11ad. Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications – amendment 3: enhancements for very high throughput in the 60 GHz band
- IEEE 802.15.3c (2009) IEEE 802.15.3c Part 15.3: wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPANs) amendment 2: millimeter-wave-based alternative physical layer extension
- Rappaport TS, Sun S, Mayzus R, Zhao H, Azar Y, Wang K, Wong GN, Schulz JK, Samimi M, Gutierrez F (2013) Millimeter wave mobile communications for 5G cellular: it will work! *IEEE Access* 1:335–349
- Rappaport TS, Heath R, Daniels RC, Murdock JN (2014) Millimeter wave wireless communications. Pearson Education, Upper Saddle River

Millimeter Wave Massive MIMO

Wei Xu¹, Yongming Huang², and Ming Xiao³

¹Southeast University, Nanjing, China

²School of Information Science and Engineering, Southeast University, Nanjing, China

³Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

[Millimeter wave large-scale MIMO](#)

Definition

Millimeter wave massive multiple-input multiple-output (MIMO) is a wireless communication technique that transmits and receives millimeter-length electromagnetic wave signals through massive antenna arrays at transmitters and receivers. Multiantenna arrays in MIMO enable simultaneous transmissions of multiple data

streams through wireless channels. In most cases, massive MIMO implicitly implies that multiple users are served simultaneously, that is, a multiuser massive MIMO. If only the transmitter deploys a multiantenna array while each receiver equips a single antenna, the terminology MIMO refers to, more formally, multiple-input single-output (MISO). If only the receiver deploys an antenna array while the transmitter equips a single antenna, the terminology MIMO refers to, more formally, single-input multiple-output (SIMO). In literature, millimeter wave massive MIMO can also represent millimeter wave massive MISO or millimeter wave massive SIMO communications.

Historical Background

The development of millimeter wave massive MIMO traces back to a natural combination of two lines of technology evolutions, i.e., millimeter wave communication and massive MIMO, in wireless communication. Millimeter wave refers to the radio spectrum with wavelength between 1 and 10 mm, corresponding to the spectrum between 30 and 300 GHz that is sometimes called the extremely high frequency (EHF) range. Millimeter wave communication was firstly experimented by J. C. Bose in 1895 realizing transmission and reception of electromagnetic waves at 60 GHz over 23 m distance (Emerson 1997). Early applications, however, emerged in radio astronomy and military nearly half a century later. Commercial applications of millimeter wave radar and communication devices started to become popular with the development of millimeter wave integrated circuits after the 1980s. Communications in millimeter wave usually operate with a much larger bandwidth that can be tens of times of the available bandwidth for wireless communications in microwave, which is able to support very high data rate communications. On the other hand, attenuation of radio waves increases exponentially in both propagation distance and the radio frequency. Millimeter wave communications thus face the challenge of high

propagation loss due to the high frequency. Moreover, compared to microwave propagation, attenuation of millimeter wave propagation also depends heavily on many other factors, e.g., atmospheric humidity, fog, and rain. Studies were conducted in investigating propagation characteristics of millimeter waves for communication system design (Rappaport et al. 2013). In 2013, millimeter wave communication was specified in the Wi-Fi standard IEEE 802.11ad in the 60 GHz spectrum to achieve high data rate transmission up to 7 Gbit/s (Wi-Fi Alliance 2016).

The invention of MIMO in wireless communication traces back to the early 1990s. The well-known MIMO transmission strategy, named BLAST, was firstly reported in Foschini (1996). Theoretical channel capacity of the MIMO channel was analyzed in Foschini and Gans (1998) in 1998. It revealed that MIMO is able to boost the channel capacity multiple times by deploying multiantenna arrays at both transmitter and receiver. Commercial applications of MIMO were successful in both Wi-Fi and 3G/4G cellular networks, e.g., by standardizing multiantenna stations and terminals with multiple antennas up to eight.

In 2010, a milestone in the development of MIMO technology was the initial introduction of massive MIMO by Dr. T. Marzetta from Bell lab (Marzetta 2010). Massive MIMO proposes to equip the device with an antenna array using a very large number of antenna elements, which is able to support many users to transmit data simultaneously and efficiently without interfering each other too much. The study in Ngo et al. (2013) discovered that the transmit power can be effectively scaled down with the number of antennas while still guaranteeing a desired transmission rate in massive MIMO. Due to its advantages, massive MIMO has been recognized as one of the most essential technology evolutions in the next-generation cellular network, namely 5G. With the deployment of a massive antenna array, however, implementation of massive MIMO in practice also faces many challenges. For instance, estimation of the large-dimensional MIMO channel is resource-hungry; deployment of a massive antenna array occupies a huge space when using

microwave spectrum; and the cost of constructing and driving a massive antenna array could be expensive (Rusek et al. 2013).

Millimeter wave massive MIMO, more than a natural combination of massive MIMO communication with millimeter wave spectrum, exhibits multiple unique features compared to conventional massive MIMO using the microwave spectrum. Besides the large amount of spectrum resource at millimeter wave band, the much shorter wavelength facilitates packing massive antennas into an array of compact size, which endues millimeter wave massive MIMO with great potential in providing ultra-high data rate communications (Xiao et al. 2017). Because millimeter wave channels are sparse in propagation paths and suffer large path-loss and blockage, the MIMO transmission and reception design of millimeter massive MIMO is substantially different from that for conventional microwave massive MIMO. In order to combat the large path-loss, power efficiency has been one of the principle considerations in system design. It becomes particularly important to form directional beams in millimeter wave MIMO by using massive antenna array (Huang et al. 2018). Transmit power can thus be mostly assembled in the directional and narrow beams to combat large path-loss and enhance communication links.

Understanding propagation characteristics of millimeter wave massive MIMO channels is crucial for both optimization of antenna array topology and design of directional beam patterns. A thorough survey was reported in Rappaport et al. (2017) on existing measurement results and modeling of millimeter wave MIMO propagations, especially in the context of their applications in 5G. A common agreement is that the channel is sparse in the angular domain, and in a large portion of application scenarios, line-of-sight propagation components play a significant part of all propagation effects. The sparsity nature, in one way, provides great potential in solving the channel estimation challenge in massive MIMO, while in the other way it limits the number of degree of freedoms (DoF) that can be achieved through the large dimensional MIMO channel.

In recent years, new signal processing techniques (Heath et al. 2016) as well as hardware-friendly architectures (Liang et al. 2014; Ayach et al. 2014) are emerging to exploit these unique features and meet corresponding challenges in millimeter wave massive MIMO. One of the well-acknowledged new solutions refers to as hybrid analog-digital precoding/combining that balances the sparsity channel feature and hardware constraints in millimeter wave MIMO circuit design (Brady et al. 2013; He et al. 2017). Design parameters in hybrid analog-digital millimeter wave massive MIMO system were analyzed in Xu et al. (2017) and Xue et al. (2017), proposing an effective DoF in performance characterization (Xue et al. 2017). Another design trend is to replace high-resolution analog-to-digital converters (ADC) in conventional massive MIMO by low-resolution ADCs in order to achieve reduction in both circuit power consumption and hardware cost for implementations. Studies, e.g., (Liu et al. 2016; Saxena et al. 2017), discovered that low-resolution ADCs cause negligible performance loss compensable by a massive antenna array in millimeter wave communications.

Key Applications

Millimeter wave massive MIMO enables ultra-high data rate communications in applications like wireless backhaul and ultra-dense small cell coverage in cellular networks. It is also becoming a highly attractive solution for realizing low latency and/or wideband communications in unmanned aerial vehicular communications and the coming intelligent vehicular communication networks (Xiao et al. 2016). Vehicular radar communication and detection is also a growing application of millimeter wave massive MIMO.

Cross-References

- [Millimeter Wave Beam Training and Tracking](#)
- [Millimeter Wave Channel Access](#)
- [Millimeter Wave Channel Modeling](#)

References

- Ayach O et al (2014) Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans Wirel Commun* 13(3):1499–1513
- Brady J, Behdad N, Sayeed A (2013) Beam-space MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements. *IEEE Trans Antennas Propag* 61(7):3814–3827
- Emerson D (1997) The work of Jagadis Chandra Bose: 100 years of mm-wave research. In: *IEEE MTT-S international microwave symposium digest*, Denver
- Foschini G (1996) Layered space-time architecture for wireless communication in a fading environment when using multiple antennas. *Labs Syst Tech J Bell* 1:41–59
- Foschini G, Gans M (1998) On limits of wireless communications in a fading environment when using multiple antennas. *Wirel Pers Commun* 6(3):311–335
- He S, Wang J, Huang Y, Ottersten B, Hong W (2017) Codebook based hybrid precoding for millimeter wave multiuser systems. *IEEE Trans Signal Process* 65(7):5289–5304
- Heath R et al (2016) An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE J Sel Top Sign Proces* 10(3):436–453
- Huang Y, Zhang J, Xiao M (2018) Constant envelope hybrid precoding for directional millimeter-wave communications. *IEEE J Sel Areas Commun*
- Liang L, Xu W, Dong X (2014) Low-complexity hybrid precoding in massive multiuser MIMO systems. *IEEE Wireless Commun Lett* 3(6):653–656
- Liu J, Xu J, Xu W, Jin S, Dong X (2016) Multiuser massive MIMO relaying with mixed-ADC receiver. *IEEE Signal Process Lett* 24(1):76–80
- Marzetta T (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wirel Commun* 9(11):3590–3600
- Ngo H, Larsson E, Marzetta T (2013) Energy and spectral efficiency of very large multiuser MIMO systems. *IEEE Trans Commun* 61(4):1436–1449
- Rappaport T et al (2013) Millimeter wave mobile communications for 5G cellular: it will work! *IEEE Access* 1:335–349
- Rappaport T et al (2017) Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models. *IEEE Trans Antennas Propag* 65(12):6213–6230
- Rusek F et al (2013) Scaling up MIMO: opportunities and challenges with very large arrays. *IEEE Signal Process Mag* 30(1):40–60
- Saxena A, Fijalkow I, Swindlehurst A (2017) Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink. *IEEE Trans Signal Process* 65(17):4624–4463
- Wi-Fi Alliance (2016) Wi-Fi Certified WiGig: Wi-Fi expands to 60 GHz. <https://www.wi-fi.org/>
- Xiao Z, Xia P, Xia X (2016) Enabling UAV cellular with millimeter-wave communication: potentials and approaches. *IEEE Commun Mag* 54(5):66–73
- Xiao M et al (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun* 35(9):1909–1935
- Xu W, Liu J, Jin S, Dong X (2017) Spectral and energy efficiency of multi-pair massive MIMO relay network with hybrid processing. *IEEE Trans Commun* 65(9):3794–3809
- Xue C, He S, Huang Y, Wu Y, Yang L (2017) An efficient beam-training scheme for the optimally designed subarray structure in mmWave LoS MIMO systems. *EURASIP J Wirel Commun Netw* 2017:31

Millimeter Wave Multiple Access

► Millimeter Wave Channel Access

Millimeter Wave NOMA

Bichai Wang¹, Linglong Dai¹, and Ming Xiao²
¹Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China

²Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Multiple access in high frequency

Definitions

Millimeter-wave NOMA refers to using non-orthogonal multiple access (NOMA) in millimeter-wave frequencies. On the one hand, millimeter-wave communications, operating from 30 to 300 GHz, provide an opportunity to meet explosive capacity demand for wireless

communications. On the other hand, multiple access technologies are necessary for supporting multiple users in wireless communication systems. Particularly, NOMA can significantly improve the spectral efficiency and connectivity density. In contrast to the conventional orthogonal multiple access (OMA) schemes realized in the time-, frequency-, code-domain or their combinations, NOMA can be realized in a new domain, i.e., the power domain. By performing superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, multiple users can be simultaneously supported at the same time-frequency-space resources, and the channel gain difference among users can be translated into multiplexing gain by superposition coding. By integrating NOMA into millimeter-wave communications, potential performance gain can be achieved.

Key Points

Multiple access technologies are necessary for supporting multiple users in mobile networks, and they have been widely investigated in the lower frequency bands. Different multiple access technologies have been utilized in practical systems, including frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA). These multiple access technologies are also applicable to millimeter-wave band, but in different flavors than in lower frequency bands due to the increased complexity caused by the greatly increased bandwidth, and the different channel characteristics in millimeter-wave bands, e.g., highly directional transmissions (Mumtaz et al. 2016).

In addition to orders-of-magnitude larger bandwidths, the smaller wavelengths at millimeter-wave allow more antennas in a same physical space, which enables massive multiple input multiple output (MIMO) to provide more multiplexing gain and beamforming gain (Swindlehurst et al. 2014). In fact, beamforming

with a large antenna array is a key characteristic of millimeter-wave communication, which is used to compensate for the high pass loss of millimeter-wave signals (Xiao et al. 2017a).

To further increase the spectral efficiency, NOMA has been recently considered in millimeter-wave communications. It has been shown that NOMA can significantly improve the spectral efficiency compared to the conventional OMA schemes (Dai et al. 2015). In Ding et al. (2017), the integration of NOMA with millimeter-wave communications has been investigated, where random steering single-beamforming was adopted, which can work only in a special case that the NOMA users are close to each other. In Xiao et al. (2017b), joint power allocation and beamforming to maximize the sum rate of a two-user millimeter-wave NOMA system was proposed, where an analog beamforming structure with a phased array was considered.

Furthermore, it is well known that in conventional MIMO systems, each antenna usually requires one dedicated radio-frequency (RF) chain to realize the fully digital signal processing (Rusek et al. 2013). In this way, the use of a very large number of antennas in millimeter-wave massive MIMO systems leads to an equally large number of RF chains, which will result in unaffordable hardware cost and energy consumption (Heath et al. 2016). To address this issue, hybrid precoding (HP) has been proposed to significantly reduce the number of required RF chains in millimeter-wave massive MIMO systems without an obvious performance loss (Rusek et al. 2013).

By using NOMA in HP-based millimeter-wave massive MIMO systems, more than one user can be supported in each beam with the aid of intra-beam superposition coding and SIC, which is essentially different from conventional millimeter-wave massive MIMO using one beam to serve only one user at the same time-frequency resources. Particularly, NOMA was applied to beamspace MIMO for the first time in Wang et al. (2017), which can be regarded as a low-complexity realization of HP, and power allocation was optimized to maximize

the achievable sum rate. In addition, NOMA was also utilized in fully connected HP architecture in Yuan et al. (2017), and digital precoding was designed by modifying the conventional block diagonalization (BD) precoding scheme. Furthermore, a more sophisticated digital precoding design, i.e., minorization-maximization (MM)-based precoding, was proposed in Zhao et al. (2017) to maximize the achievable sum rate. Besides, power allocation was optimized in Hao et al. (2017) to maximize the energy efficiency of millimeter-wave massive MIMO-NOMA systems, and an iterative algorithm was proposed to obtain the optimal power allocation.

Note that the highly direction feature of millimeter-wave propagation makes the users' channels (along the same or similar direction) highly correlated, which facilitates the integration of NOMA in millimeter-wave communication. Therefore, considering the harmony between the millimeter-wave channel characteristics and the principle of NOMA, the use of NOMA for millimeter-wave communications is a very promising research direction deserves further investigations.

Cross-References

- [Lens Antenna Array](#)
- [Millimeter Wave Massive MIMO](#)

References

- Dai L, Wang B, Yuan Y, Han S, Chih-Lin I, Wang Z (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun Mag* 53(9):74–81
- Ding Z, Fan P, Poor HV (2017) Random beamforming in millimeter-wave NOMA networks. *IEEE Access* 5:7667–7681
- Hao W, Zeng M, Chu Z, Yang S (2017) Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access. *IEEE Wireless Commun Lett* 6(6):782–785
- Heath RW, Gonzalez-Prelcic N, Rangan S, Roh W, Sayeed AM (2016) An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE J Sel Top Sign Proces* 10(3):436–453
- Mumtaz S, Rodriguez J, Dai L (2016) *MmWave massive MIMO: a paradigm for 5G*. Academic Press, London/San Diego
- Rusek F, Persson D, Lau BK, Larsson EG, Marzetta TL, Edfors O, Tufvesson F (2013) Scaling up MIMO: opportunities and challenges with very large arrays. *IEEE Signal Process Mag* 30(1):40–60
- Swindlehurst AL, Ayanoglu E, Heydari P, Capolino F (2014) Millimeter-wave massive MIMO: the next wireless revolution? *IEEE Commun Mag* 52(9):56–62
- Wang B, Dai L, Wang Z, Ge N, Zhou S (2017) Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array. *IEEE J Sel Areas Commun* 35(10):2370–2382
- Xiao Z, Dai L, Ding Z, Choi J, Xia P (2017a) Millimeter-wave communication with non-orthogonal multiple access for 5G. *arXiv preprint arXiv:170907980*
- Xiao Z, Zhu L, Choi J, Xia P, Xia XG (2017b) Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter-wave communications. *arXiv preprint arXiv:171101380*
- Yuan W, Kalokidou V, Armour SMD, Doufexi A, Beach A (2017) Application of non-orthogonal multiplexing to mmWave multi-user systems. In: *IEEE vehicular technology conference (IEEE VTC'17 Spring)*, IEEE, pp 1–5
- Zhao Y, Xu W, Jin S (2017) An minorization-maximization based hybrid precoding in NOMA-mMIMO. In: *IEEE international conference on wireless communications and signal processing (IEEE WCSP'17)*, IEEE, pp 1–6

Millimeter Wave Physical Layer Security

- [Millimeter Wave Security](#)

Millimeter Wave Security

Pei Zhou, Qing Xue, and Xuming Fang
Southwest Jiaotong University, Chengdu, China

Synonyms

[Millimeter Wave Communication Security](#); [Millimeter Wave Physical Layer Security](#)

Definition

Different from the traditional communication security methods which provide data security by cryptographic techniques, physical layer security methods guarantee secure wireless transmissions by exploiting the imperfections of the communications medium (Yang et al. 2015). With physical layer security methods, communication security can be achieved by degrading the quality of signal reception at potential eavesdroppers and therefore preventing them from acquiring confidential information from the received signal. However, compared to microwave communications, severe path loss in millimeter wave (mmWave) makes it hard to support long-distance transmissions. Thanks to the short wavelength of mmWave, it is possible to employ a large number of antennas in a small size terminal. With the help of the large number of antennas, directional beamforming and multiple-antenna technologies, etc. become the powerful tools for achieving long-distance communications and enhancing the physical layer security in mmWave networks. With the degrees of freedom provided by multiple antennas (Wang and Wang 2016), better communication security can be achieved in mmWave communications.

Key Points

The conventional communication security can be achieved by ensuring all involved entities load proper and authenticated cryptographic information. However, the physical layer security does not consider about how those security protocols are executed and does not require to implement any extra security schemes or algorithms on other layers above the physical layer (Liu et al. 2017). There are many techniques that can be used for improving the security of mmWave, such as directional transmission, antenna subset modulation, and directional modulation.

Directional Transmission: Different from microwave networks, directional transmission

techniques should be used in mmWave networks to achieve better performance. Thus, the signal or interference power in mmWave communications is highly directional and closely related to critical parameters, such as transmission distance, transmit power, offset angle of departure/arrival, and beamwidth. They all have impact on the performance of mmWave communications due to the inherent directivity. If one or more of these parameters do not reach the predefined thresholds, the directional communication in mmWave networks will be damaged. Therefore, it is hard for eavesdroppers to acquire confidential information from the received signal. Meanwhile, physical layer security could naturally gain the benefit from the directivity.

Antenna Subset Modulation: With multi-antenna techniques, by using random antenna subsets and performing analog precoding with antenna selection, instead of using all antennas for beamforming, a random set of antennas are co-phased to transmit the information symbol to the desired receiver, while the rest of the antennas are used to randomize the far field radiation pattern at non-desired receiver directions. The indices of these antennas are randomized in every symbol transmission. Thus, coherent transmission to the desired receiver and a noise-like signal that jams potential eavesdroppers can be achieved (Eltayeb et al. 2017). Similarly, another antenna subset modulation method to realize physical layer security of mmWave is to modulate the radiation pattern at the symbol rate by driving only a subset of antennas in the array. This results in a directional radiation pattern that projects a sharply defined constellation in the desired direction and expanded further randomized constellation in other directions (Valliappan et al. 2013). The way to implement antenna subset modulation is randomly selecting an antenna subset for every symbol. The symbol modulation for a desired receiver along the main lobe will not be affected by randomly switching antenna subsets.

Directional Modulation: Several directional modulation approaches have been proposed that leverage multiple transmit antennas including near field antenna-level modulation,

switched antenna phased array transmitters, and spatial keying transmission techniques such as spatial modulation and space shift keying to achieve enhanced security (Valliappan et al. 2013). A hybrid multiple-input multiple-output (MIMO) phased array time-modulated directional modulation scheme was proposed to improve the physical layer security of mmWave communications (Wang and Zheng 2018). Because the hybrid MIMO phased array can take advantage of the spatial diversity of MIMO, the main advantage of phased array in coherent directional gain will not be sacrificed. Then, the transmit antenna arrays will be divided into multiple subarrays, and each subarray can be used to form a directional beam. All subarrays are jointly combined to work as an MIMO for higher angular resolution or multiuser communications. Therefore, even the eavesdropper's position is unknown, physical layer security can also be realized by applying a time-modulated directional modulation scheme.

In addition, heterogeneous networks (Het-Nets) architecture has attracted a lot of research interests recently. To provide physical layer security in the new HetNet architecture, dense small cells are deployed to bring ubiquitous inter-tier and intra-tier interference (Zhu et al. 2017). Such interference can be utilized for confounding the eavesdroppers and achieving secure communications at the physical layer as mentioned in *antenna subset modulation*.

We can see from the above physical layer security methods that they do not rely on upper-layer data encryption or secret keys. Thus, the processing overhead and additional communication overhead can be reduced. Physical layer security of mmWave communications is a promising research topic for mmWave security.

Key Applications

In order to provide ultrahigh-speed wireless transmissions for future communication systems,

mmWave will play an important role in wireless local area networks (WLANs), fifth generation (5G) mobile communication system, vehicular communication system, and so on. With the increasing demand of privacy and security of every user, communication security is critical to all the above communication systems. mmWave security technologies can be used in any mmWave networks to provide the desired users secure communications and prevent the information from being obtained by potential eavesdroppers.

Cross-References

- Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks
- Millimeter Wave Massive MIMO

References

- Eltayeb ME, Choi J, Al-Naffouri TY, Heath RW (2017) Enhancing secrecy with multiantenna transmission in millimeter wave vehicular communication systems. *IEEE Trans Veh Technol* 66(9):8139–8151
- Liu Y, Chen HH, Wang L (2017) Physical layer security for next generation wireless networks: theories, technologies, and challenges. *IEEE Commun Surv Tut* 19(1):347–376
- Valliappan N, Lozano A, Heath RW (2013) Antenna subset modulation for secure millimeter-wave wireless communication. *IEEE Trans Commun* 61(8):3231–3245
- Wang C, Wang HM (2016) Physical layer security in millimeter wave cellular networks. *IEEE Trans Wirel Commun* 15(8):5569–5585
- Wang WQ, Zheng Z (2018) Hybrid MIMO and phased-array directional modulation for physical layer security in mmWave wireless communications. *IEEE J Sel Areas Commun*. <https://doi.org/10.1109/JSAC.2018.2825138>
- Yang N, Wang L, Geraci G, Elkashlan M, Yuan J, Di Renzo M (2015) Safeguarding 5G wireless communication networks using physical layer security. *IEEE Commun Mag* 53(4):20–27
- Zhu Y, Wang L, Wong KK, Heath RW (2017) Secure communications in millimeter wave ad hoc networks. *IEEE Trans Wirel Commun* 16(5):3205–3217

Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique

Xinyu Gao¹, Linglong Dai², Yongming Huang³, and Ming Xiao⁴

¹Department of Electronic Engineering, Tsinghua University, Beijing, China

²Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China

³School of Information Science and Engineering, Southeast University, Nanjing, China

⁴Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

High-frequency MIMO; Millimeter-wave massive MIMO

Definitions

Multiple-input multiple-output (MIMO) is a technique that employs multiple transmit and receive antennas to send and receive more than one data stream simultaneously over the same time/frequency resource. Millimeter-wave (mmWave) MIMO is the MIMO technique operated at mmWave frequencies (i.e., 30–300 GHz).

Historical Background

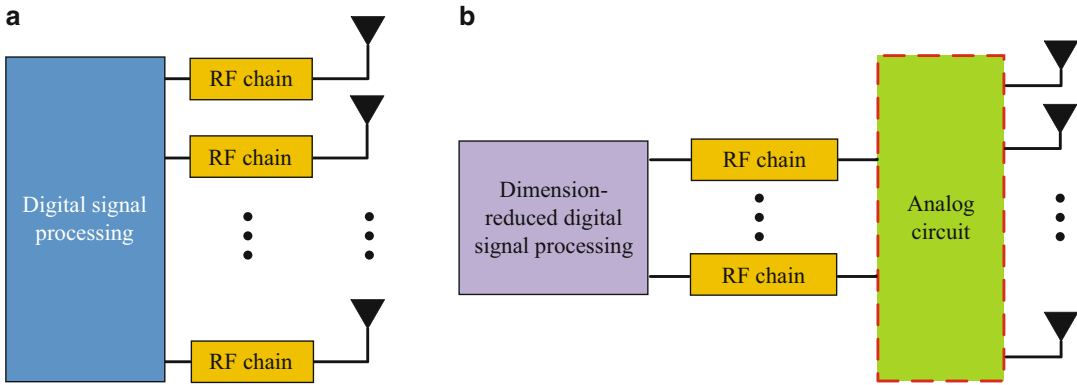
The history of MIMO can be traced back to 1993, when Arogyaswami Paulraj and Thomas Kailath proposed to “split a high-rate signal into several low-rate signals to be transmitted from spatially separated transmitters and recovered by

the receive antenna array based on differences in directions-of-arrival” (Li et al. 2010). After that, this technique has attracted great interest from both the academic and industry communities and obtained a huge development. Currently, MIMO technique has been widely used in wireless communication systems working at sub-6 GHz frequencies, such as 4G cellular systems (Li et al. 2010). By employing multiple antennas, MIMO can achieve multiplexing gain with reduced interference to improve the spectrum efficiency.

However, the MIMO technique in current 4G cellular systems only employs a small number of transmit antennas and receive antennas due to the limited physical space at the transmitter and receiver. As a result, the improvement in the spectrum efficiency is still limited, which cannot meet the one thousand times increase in data traffic predicted for further 5G cellular systems (Mumtaz et al. 2016). To solve this problem, one possible way is to extend the MIMO technique at sub-6 GHz frequencies to mmWave frequencies (Mumtaz et al. 2016). On one hand, mmWave band can provide more than 2 GHz bandwidth for communication, which is much wider than the 20 MHz bandwidth in current 4G cellular systems. On the other hand, the short wavelengths associated with mmWave frequencies enable a large antenna array to be packed in a small physical space, which means that MIMO with a large antenna array can be easily employed at mmWave frequencies to considerably improve the spectrum efficiency.

Foundations

The mmWave MIMO technique will be quite different from the one at sub-6 GHz frequencies due to the different channel characteristics and additional hardware constraints found at mmWave frequencies (Gao et al. 2018). The conventional MIMO architecture is fully digital as shown in Fig. 1a, where each antenna requires one dedicated RF chain (including power amplifier, low-noise amplifier, data converter, mixer, and so on) to perform digital signal processing in the baseband (Li et al. 2010). For mmWave MIMO,



Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique, Fig. 1 MIMO architectures: (a) fully digital architecture; (b) hybrid analog and digital architecture

this architecture will suffer from unaffordable hardware cost and power consumption, since (1) the number of antennas is usually large due to the short wavelengths (e.g., 256 antennas at mmWave frequencies instead of 8 antennas at sub-6 GHz frequencies) (Mumtaz et al. 2016) and (2) the power consumption of RF chain is high due to the increased sampling rate (e.g., 250 mW/RF chain at mmWave frequencies, compared with 30 mW/RF chain at sub-6 GHz frequencies) (Gao et al. 2018).

To this end, the hybrid analog and digital architecture is proposed for mmWave MIMO (Gao et al. 2018), as shown in Fig. 1b. This architecture divides the conventional digital signal processing of large size into two parts, i.e., a large-size analog signal processing (realized by analog circuit) and a dimension-reduced digital signal processing (requiring a small number of RF chains). The smallest number of required RF chains in hybrid architecture can equal the number of data streams for communication, which is usually much smaller than the number of antennas. In this way, the number of RF chains can be significantly reduced, leading to quite low energy consumption. On the other hand, due to the limited number of effective scatterers at mmWave frequencies, the mmWave MIMO channel matrix is usually low-rank (Gao et al. 2018), and the maximum number of data streams that can be simultaneously transmitted in such low-rank channel is small. Therefore,

as long as the number of RF chains is larger than the rank of channel matrix, the dimension-reduced digital signal processing is still able to fully achieve the multiplexing gain. As a result, hybrid architecture can achieve a better trade-off between the hardware cost/energy consumption and the sum-rate performance (Gao et al. 2018).

Architecture for mmWave MIMO

The analog circuit of mmWave MIMO with hybrid architecture can be implemented by different hardware networks (Gao et al. 2018; Méndez-Rial et al. 2016) as listed below.

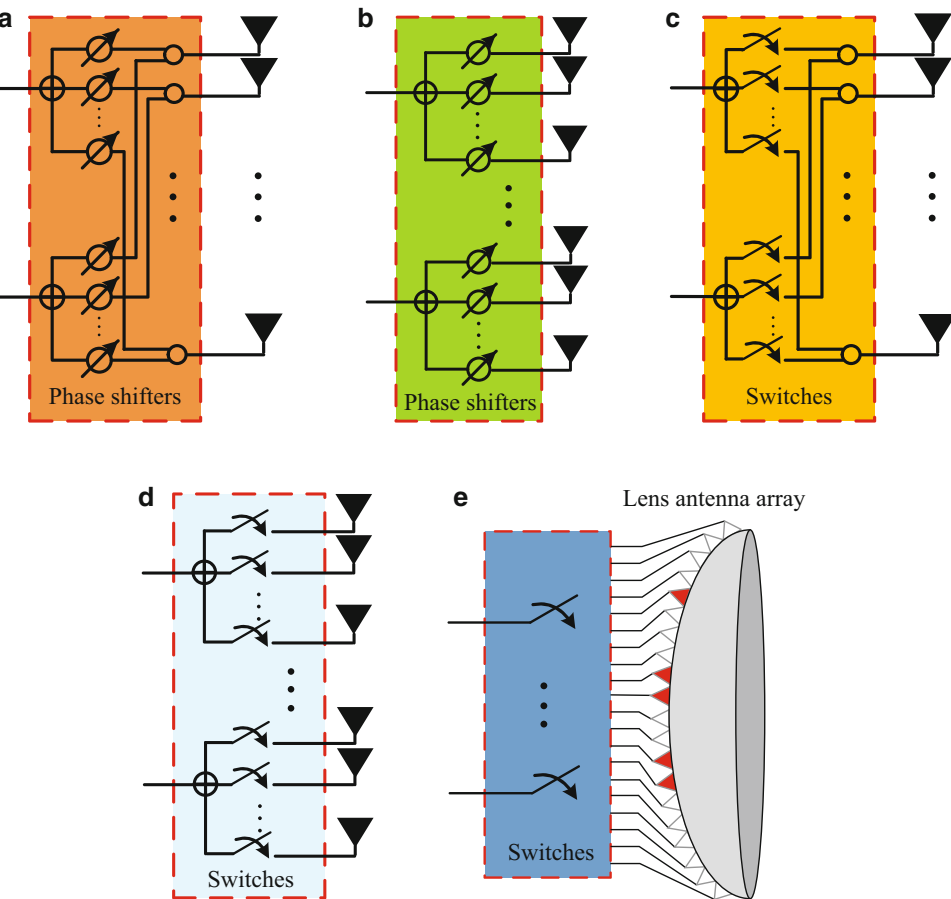
- (N1) *Fully connected network with phase shifters*: In this network, each RF chain is connected to all antennas via phase shifters, as shown in Fig. 2a. For each RF chain, full array gain can be achieved by adjusting the phases of transmitted signals on all antennas to provide high array gain. By employing such network, all elements of the analog precoder/combiner have the same fixed amplitude.
- (N2) *Sub-connected network with phase shifters*: In this network, each RF chain is only connected to a sub-antenna array via phase shifters, as shown in Fig. 2b. Due to this limitation, for each RF chain, only the transmitted signals on a subset

of antennas can be adjusted. Therefore, compared with N1, the achieved array gain in this network is reduced. However, this network is preferred in practice, since the number of phase shifters required in this network is significantly reduced to save energy consumption. This network imposes two hardware constraints: (i) the analog precoder/combiner should be a block diagonal matrix; (ii) all the nonzero elements of the analog precoder/combiner have the same fixed amplitude.

(N3) *Fully connected network with switches:* In this network, each RF chain is connected to all antennas similar to N1 but via low-

cost switches instead of phase shifters as shown in Fig. 2c. Each RF chain selects several antennas for communication, and the transmitted signals cannot be adjusted. Therefore, this network will suffer from serious degradation of array gain. However, as the phase shifters are replaced by simple switches, the hardware cost and energy consumption can be considerably reduced. In this network, all the elements of the analog precoder/combiner can be only selected from the set {0, 1}.

(N4) *Sub-connected network with switches:* In this network, each RF chain is connected to a sub-antenna array like N2 with phase



Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique, Fig. 2 Analog circuit with different networks: (a) Fully connected network with phase shifters (N1); (b) Sub-connected network with

phase shifters (N2); (c) Fully connected network with switches; (d) Sub-connected network with switches; (e) Lens antenna array

shifters being replaced by switches, as shown in Fig. 2d. Each RF chain can only select a subset of antennas instead of the whole antenna array to transmit signals. Obviously, the achieved array gain is further reduced compared with N3. However, it also enjoys the lowest hardware cost and energy consumption, as it requires only a small number of switches without the need of splitters. The hardware constraints induced by this network are similar to those of N2, i.e., the analog precoder/combiner should be a block diagonal matrix, but its nonzero elements can be only selected from $\{0, 1\}$.

- (N5) *Lens antenna array*: An alternative quite different from the four networks discussed above is to utilize lens antenna array, as shown in Fig. 2e. The lens antenna array (a feed antenna array placed beneath the lens) (Brady et al. 2013) can realize the functions of signal emitting and phase shifting simultaneously. It can concentrate the signals from different propagation directions (beams) on different feed antennas. As the scattering at mmWave frequencies is not rich (Gao et al. 2018), the channel power will be concerted on only a small number of beams. Therefore, the simple selecting network with switches can be used to significantly reduce the MIMO dimension as well as the number of RF chains without obvious performance loss. Moreover, since all the phase shifters and splitters are replaced by one simple lens, the hardware cost and energy consumption of this network is also considerably low. Mathematically, the lens antenna array plays the role of a discrete Fourier transform (DFT). Therefore, in this network, each column of the analog precoder/combiner restricts to an DFT column.

Signal Processing for mmWave MIMO

The signal processing (including precoding/combining, channel estimation, and so on) for mmWave MIMO is also different from the

one for conventional MIMO since the fully digital architecture is replaced by the hybrid architecture.

We first discuss how to design the hybrid precoding (includes digital precoder and analog precoder) and combining (includes digital combiner and analog combiner) for mmWave MIMO with hybrid architecture. In general, the design target is to maximize the achievable sum rate. However, obtaining the optimal solution is not a trivial task, and the main difficulties are twofold. Firstly, the designs of digital precoder/combiner and analog precoder/combiner are coupled, which makes the optimization problem non-convex. Secondly, as described above, the analog circuit of hybrid architecture imposes non-convex hardware constraints on the analog precoder/combiner. To alleviate these two challenges, one feasible way is to decompose the original optimization problem into several subproblems, and each subproblem is approximated as a convex one and then solved by efficient convex optimization algorithms.

For example, in El Ayach et al. (2014), a spatially sparse precoding scheme is proposed for N1. It first assumes that the hybrid combiner at the receiver is ideal without any hardware constraint. Then, it approximates the sum-rate optimization problem as the one minimizing the distance between the optimal fully digital precoder without any constraint and the hybrid precoder. Then, a variant of the orthogonal matching pursuit (OMP) algorithm is developed to obtain the near-optimal hybrid precoder. Finally, the hybrid combiner can be designed in a similar way based on the effective channel. For N2, a successive interference cancelation (SIC)-based precoding scheme is proposed (Gao et al. 2016). Similar to the spatially sparse precoding scheme, it first decouples the design of precoder and combiner. Then, it decomposes the total achievable sum-rate optimization problem with non-convex constraints into a series of simple sub-rate optimization problems, each of which only considers one sub-antenna array. After that, it obtains the optimal hybrid precoding vector for each sub-antenna array, which is sufficiently close to the unconstrained optimal solution. Then, borrowing the concept of SIC for multiuser detec-

tion, the achievable sub-rate of each sub-antenna array is optimized one by one until the last sub-antenna array is considered. After that, the hybrid combiner can be obtained in a similar way based on the effective channel. Besides these two schemes, some other hybrid precoding and combining schemes designed for N1-N5 can be further found in Méndez-Rial et al. (2016).

The optimal performance of hybrid precoding and combining can be only achieved with perfect channel state information (CSI). However, for mmWave MIMO with hybrid architecture, estimating the complete CSI is not a trivial task (Alkhateeb et al. 2014). Firstly, due to the lack of antenna gain before the establishment of the transmission link, the SNR for channel estimation in mmWave MIMO is quite low. Secondly, the number of RF chains in hybrid architecture is usually much smaller than the number of antennas, so we cannot simultaneously obtain the sampled signals on all antennas. As a result, the traditional channel estimation schemes requiring the sampled signals on all antennas will involve unaffordable pilot overhead in mmWave MIMO. To solve this problem, two dominant categories of channel estimation schemes have been proposed.

The key idea of the first category is to reduce the dimension of channel estimation problem. For example, in Hogan and Sayeed (2016), a two-step channel estimation scheme is proposed for N5. In the first step, it performs the beam training between the transmitter and receiver to obtain the analog precoder and analog combiner. In the second step, the effective channel matrix in the analog domain is estimated by classical algorithms, such as least squares (LS). Note that the size of the effective channel matrix is much smaller than that of the original channel matrix. Therefore, the pilot overhead in the second step is quite low. The second category of channel estimation schemes is to exploit the sparsity of mmWave MIMO channels. Instead of estimating the effective channel matrix of small size, it can directly obtain the complete channel matrix with low pilot overhead. For example, in Alkhateeb et al. (2014), an adaptive channel estimation scheme is proposed for N1. It divides the total channel esti-

mation problem into several subproblems, each of which only considers one channel path. For each channel path, it first starts with a coarse angle of arrival (AoA)/angle of departure (AoD) grids and determines the AoA and AoD of this path belonging to which angle range by employing OMP algorithm. Then, the narrowed direction grids are used, and the direction of this path is further refined. Besides these two schemes, some other channel estimation schemes designed for N1-N5 can be further found in Méndez-Rial et al. (2016).

Key Applications

MmWave MIMO has been considered as a promising technique for further wireless communication systems. The applications of mmWave MIMO are immense. For example, mmWave MIMO could be used to enable high-rate low-latency data transmission for 5G cellular systems. In addition, with the recent excitement related to connected and autonomous vehicles, mmWave MIMO can play an important role in providing high data rate connections between cars. Finally, mmWave MIMO is also of interest for high-speed wearable networks that connect cell phone, smart watch, augmented reality glasses, and virtual reality headsets.

Future Directions

There are still some open issues on mmWave MIMO. For example, most of the existing signal processing for mmWave MIMO is designed under the narrowband and time-invariant channels. However, due to the large bandwidth and the high frequency, the mmWave MIMO channels are more likely to be broadband and time-varying, which incurs new challenges. Take the hybrid precoding for example. Broadband means that the analog precoder cannot be adaptively adjusted according to the frequency, leading to more difficulties in signal processing design, while time-varying means that we need to re-estimate the channel and re-compute the hybrid precoding frequently, leading to high pilot

overhead and computational complexity. Therefore, designing signal processing for mmWave under broadband time-varying channels will be an urging problem to solve.

Cross-References

- [Hybrid Precoding](#)
- [Massive MIMO Channel Estimation](#)

References

- Alkhateeb A, El Ayach O, Leus G, Heath RW (2014) Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE J Sel Topics Signal Process* 8(5):831–846
- Brady J, Behdad N, Sayeed A (2013) Beam-space MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements. *IEEE Trans Antennas Propag* 61(7):3814–3827
- El Ayach O, Rajagopal S, Abu-Surra S, Pi Z, Heath RW (2014) Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans Wireless Commun* 13(3):1499–1513
- Gao X, Dai L, Han S, I CL, Heath RW (2016) Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays. *IEEE J Sel Areas Commun* 34(4):998–1009
- Gao X, Dai L, Sayeed A (2018) Low RF-complexity technologies to enable millimeter-wave mimo with large antenna array for 5G wireless communications. *IEEE Commun Mag* 56(4):211–217
- Hogan J, Sayeed A (2016) Beam selection for performance-complexity optimization in high-dimension MIMO systems. In: *Conference on information science and systems*, pp 337–342
- Li Q, Li G, Lee W, Lee MI, Mazzaresse D, Clerckx B, Li Z (2010) MIMO techniques in WiMAX and LTE: a feature overview. *IEEE Commun Mag* 48(5):86–92
- Méndez-Rial R, Rusu C, González-Prelcic N, Alkhateeb A, Heath RW (2016) Hybrid MIMO architectures for millimeter wave communications: phase shifters or switches? *IEEE Access* 4:247–267
- Mumtaz S, Rodriguez J, Dai L (2016) *MmWave massive MIMO: a paradigm for 5G*. Academic Press, Elsevier

Millimeter-Wave Communications

- [Per-Beam Synchronization for Millimeter-Wave Massive MIMO](#)

Millimeter-Wave Massive MIMO

- [Millimeter-Wave \(mmWave\) Multiple-Input Multiple-Output \(MIMO\) Technique](#)

MIMO

- [Index Modulation](#)

MIMO Detection

Lin Bai¹, Tian Li², and Quan Yu³

¹School of Electronic and Information Engineering, Beijing Laboratory for General Aviation Technology (Beihang University), Beijing, China

²The 54th Research Institute of CETC, Shijiazhuang Hebei, China

³School of Electronic and Information Engineering, Beihang University, Beijing, China

Synonyms

[MIMO signal demodulation](#); [MIMO signal equalization](#)

Definitions

MIMO detection is to jointly detect multiple signals at the receiver, where the signals are transmitted through a wireless channel.

Historical Background

MIMO is an important technology in B3G/4G wireless communication systems, where both transmitters and receivers are equipped with antenna arrays. By exploiting the spatial diversity

benefitted from the multiple antennas, MIMO was first developed by Marconi in 1908 to combat channel fading. Then, the staff in Bell Laboratory extended this technology in different transmission environment. In 1995, Telatar studied the channel capacity of an $N_r \times N_t$ point-to-point MIMO system under Rayleigh fading condition and proofed that the achievable rate increases linearly with $\min(N_r, N_t)$ (Telatar 1999). This improved capacity is then regarded as the spatial multiplexing gain. In order to fully exploit the spatial multiplexing gain provided by MIMO systems, different Bell-Labs layered space time (BLAST) architectures were developed from 1996 to 1998 (Foschini 1996, 1999; Wolniansky 1998). As an opposite performance metric, spatial diversity gain is used to judge the transmission quality of a MIMO system. To maximize the diversity gain, Alamouti proposed a space-time block code (STBC) in 1998 (Alamouti 1998). With the STBC method, a full diversity gain, i.e., $N_t N_r$, can be obtained. In general, BLAST architectures using different MIMO detection schemes are always applied to achieve a high spectral efficiency without losing too much diversity gain. Thus, MIMO detection has been widely considered a key technology in designing MIMO systems.

Foundations

A typical MIMO system is shown in Fig. 1, where the transmission model is given by:

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n}. \quad (1)$$

Here, $\mathbf{H}$, $\mathbf{s}$, and $\mathbf{n}$ denote the channel matrix, transmitted signal vector, and background noise, respectively. MIMO detection is to recover the

transmitted signal at the receiver, where the estimated signal is denoted by $\tilde{\mathbf{s}}$ in Fig. 1.

MIMO detection in uncoded systems In general, to obtain an optimal bit-error rate (BER) performance with a full diversity gain in uncoded MIMO systems, maximum likelihood (ML) detection method can be carried out by employing exhaustive search, where the estimated signal vector is derived as $\tilde{\mathbf{s}} = \arg \min_{\mathbf{s} \in \mathcal{S}^{N_t}} \|\mathbf{r} - \mathbf{H}\mathbf{s}\|^2$. Here, $\mathcal{S}$ denotes the symbol alphabet. Due to the involved constellation searching, the computational complexity of the ML method grows exponentially with the number of transmit antennas, which makes it infeasible to be applied in a real MIMO system. Then, several low-complexity detection schemes have been widely studied in literatures (Wolniansky 1998; Bai 2012; Yao 2002; Liu 2011). Among the proposed sub-optimal methods, linear detectors are extensively applied in practice, such as zero-forcing (ZF) and minimum mean square error (MMSE) linear filters (Bai 2012):

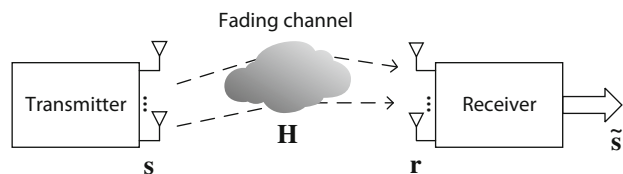
- *ZF detectors.* The signal received by a linear detector is filtered by a well-designed matrix which enables the transmitted multiple signal to be detected separately without involving interferences. For the ZF method, the filter matrix can be designed by inverting the channel matrix, i.e., $\mathbf{W}_{ZF} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H$. Then, the estimated signal vector is given by:

$$\begin{aligned} \tilde{\mathbf{s}} &= \mathbf{W}_{ZF} \mathbf{r} \\ &= \mathbf{s} + (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \mathbf{n}. \end{aligned} \quad (2)$$

However, as the equivalent noise becomes $(\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \mathbf{n}$ in (2), the noise power would be amplified if the channel matrix is not orthogonal. Furthermore, the ZF detector

MIMO Detection, Fig. 1

An illustration of a MIMO system



cannot offer a satisfied BER performance when the signal-to-noise ratio is low.

- **MMSE detectors.** In order to alleviate the co-channel interference while restraining the impact of background noise at the same time, the MMSE method was proposed. In an MMSE detector, the filter matrix is developed on the criterion of minimizing the estimation error, i.e., $\mathbf{W}_{MMSE} = \arg \min \mathbb{E}[\|\mathbf{s} - \mathbf{W}\mathbf{r}\|^2] = \mathbf{H}(\mathbf{H}^H\mathbf{H} + N_0/E_s\mathbf{I})^{-1}$. Here, N_0 and E_s denote the power of the noise and average signal, respectively. Compared with the ZF method, the filter matrix in MMSE approach can be designed by finding a balance between interference cancelation and noise amplification without incurring too much complexity.

Although the MMSE detector can improve the BER performance of ZF method, the noise will still be enhanced when filtering the co-channel interference. In Wolniansky (1998), a QR factorization-based successive interference cancelation (SIC) was proposed, where the detection of each signal can be carried out after removing the other estimated signals. Specifically, by multiplying $\mathbf{Q}^H$, where $\mathbf{Q}$ is the unitary component in the QR factorization of $\mathbf{H}$, the signal vector at the receiver is derived as

$$\begin{aligned}\mathbf{v} &= \mathbf{Q}^H\mathbf{r} \\ &= \mathbf{Q}^H\mathbf{Q}\mathbf{R}\mathbf{s} + \mathbf{Q}^H\mathbf{n} \\ &= \mathbf{R}\mathbf{s} + \mathbf{Q}^H\mathbf{n}.\end{aligned}\quad (3)$$

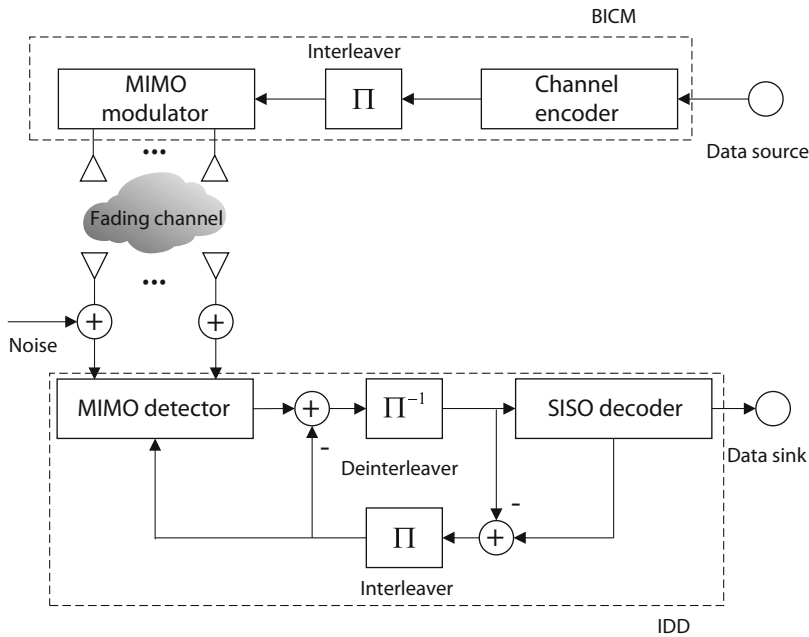
In (3), $\mathbf{R} = \mathbf{Q}^H\mathbf{H}$. Since $\mathbf{R}$ is upper triangular, signals in $\mathbf{s}$ can be estimated from the N_t th layer to the first layer in a successive manner. In summary, the SIC detector outperforms the above linear detectors because the noise would not be amplified by multiplying a unitary matrix.

In order to achieve the same diversity provided by the ML method, Yao (2002) proposed a lattice reduction-aided MIMO detection method. From the point of view of lattice decoding, MIMO detection is to find the closest vector point in a lattice generated by the basis vectors of the

channel matrix. Applying the Lenstra-Lenstra-Lovász algorithm (Lenstra 1982), a basis can be transformed into a nearly orthogonal one. Thus, the decoding radius is increased which helps to find the correct transmitted signal vector. Inspired by lattice decoding, Liu (2011) applied the Klein sampling algorithm Klein (2000) in MIMO detections. In this scheme, the searching area can be further extended by involving a randomized rounding instead of the conventional one.

MIMO detection in coded systems To improve both the channel capacity and BER performance, coded MIMO systems are widely considered, where the bit-interleaved coded modulation (BICM) is usually adopted at the transmitter (Caire 1998). At the receiver side, the iterative detection and decoding (IDD) architecture (Hochwald 2003) based on turbo principle is employed, which is shown in Fig. 2. With a soft-input soft-output (SISO) detector and a SISO decoder, the extrinsic information, i.e., log-likelihood ratio of each bit, is exchanged iteratively between them to obtain a good trade-off of the BER performance and complexity.

For MIMO detectors in IDD, the maximum a posteriori probability (MAP) method can provide an optimal BER performance. Unfortunately, since the MAP detector adopts the same exhaustive search with the ML method, the computational complexity becomes $O(2^{N_t}N_t^2)$ which also makes it infeasible to be implemented in a real MIMO system. In Wang (1999), an MMSE-based soft cancelation (SC) method was proposed to provide a sub-optimal performance with relatively low computational complexity. Since the MMSE-SC is developed on a symbol-wise linear filter, the offered BER performance is always taken as a benchmark when designing low-complexity sub-optimal methods. To improve the performance without incurring too much complexity, a bit-wise MMSE filter was studied in Li (2013). By exploring the distribution of the a posteriori probability of the transmitted signal vector, (Bai 2013) applied the Klein sampling in IDD and proposed an a priori information-based sampler. After that, Bai



MIMO Detection, Fig. 2 An illustration of a MIMO-BICM system

(2016) considered a large-scale antenna scenario and developed a Klein sampling-aided Markov chain Monte Carlo (MCMC) detector. In this scheme, the detection is carried out in a block-wise manner using the MCMC approach, while each block can be drawn with the Klein sampler.

Key Applications

MIMO detection is fundamental in wireless communication systems where the transmitter and the receiver are equipped with multiple antennas. Another typical application scenario can be found in multiuser detection.

Cross-References

- [5G Wireless](#)
- [Key Technologies in 4G/LTE Network](#)
- [Millimeter Wave Massive MIMO](#)

References

- Alamouti SM (1998) A simple transmit diversity technique for wireless communications. *IEEE J Sel Areas Commun* 16(8):1451–1458
- Bai L, Choi J (2012) Low complexity MIMO detection. Springer, New York
- Bai L, Choi J (2013) Lattice reduction-based MIMO iterative receiver using randomized sampling. *IEEE Trans Wireless Commun* 12(5):2160–2170
- Bai L, Li T, Liu J et al (2016) Large-Scale MIMO detection using MCMC approach with blockwise sampling. *IEEE Trans Commun* 64(9):3697–3707
- Caire G, Taricco G, Biglieri E (1998) Bit-interleaved coded modulation. *IEEE Trans Inf Theory* 44(3):927–946
- Foschini GJ (1996) Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Tech J* 1(2):41–59
- Foschini GJ, Golden GD, Valenzuela RA et al (1999) Simplified processing for high spectral efficiency wireless communication employing multi-element arrays. *IEEE J Sel Areas Commun* 17(4):1841–1853
- Hochwald BM, Brink ST (2003) Achieving near-capacity on a multiple-antenna channel. *IEEE Trans Commun* 51(3):389–399

- Klein P (2000) Finding the closest lattice vector when it's unusually close. In: Proceedings of ACM-SIAM symposium on discrete algorithms, San Francisco, pp 937–941
- Li Q, Zhang J, Bai L et al (2013) Lattice reduction-based approximate MAP detection with bit-wise combining and integer perturbed list generation. *IEEE Trans Commun* 61(8):3259–3269
- Liu S, Ling C, Stehle D (2011) Decoding by sampling: a randomized lattice algorithm for bounded distance decoding. *IEEE Trans Inf Theory* 57(9):5933–5945
- Lenstra AK, Lenstra HWJ, Lovász L (1982) Factoring polynomials with rational coefficients. *Math Ann* 261(4):515–534
- Telatar E (1999) Capacity of multi-antenna Gaussian channels. *Europ Trans Telecommun* 10(6):585–595
- Wang X, Poor HV (1999) Iterative (turbo) soft interference cancellation and decoding for coded CDMA. *IEEE Trans Commun* 47(7):1046–1061
- Wolniansky PW, Foschini GJ, Golden GD et al (1998) V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel. In: Proceedings of 1998 IEEE ISSSE, Pisa, pp 295–300
- Yao H, Wornell GW (2002) Lattice-reduction-aided detectors for MIMO communication systems. In: Proceedings of 2002 IEEE GLOBECOM, pp 424–428

MIMO Relay

Yue Rong

School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA, Australia

Synonyms

[MIMO relay channel](#); [MIMO relay network](#)

Definitions

Multiple-input multiple-output (MIMO) relay refers to a cooperative communication technology employing relay nodes, when one or more nodes of a relay system have multiple transmit/receive dimensions. MIMO relay communications can improve the reliability and extend the coverage of communication systems.

Historical Background

A relay channel, also known as three-terminal communication channel, was first investigated in van der Meulen (1971). The capacity of this channel was studied in 1979 (Cover and El Gamal 1979). During 1980s and 1990s, further information theoretic results on the relay channel were discovered (Vanroose and van der Meulen 1992). Since the turn of the century, the relay channel has attracted a renewed interest due to the boom of applications of wireless communications (Sendonaris et al. 2003).

When multiple antennas are deployed at one or more nodes of the relay system, we call such relay system a MIMO relay channel. The achievable rate and capacity upper bound of a MIMO relay channel have been studied in Wang et al. (2005). A diversity-multiplexing trade-off of multi-antenna cooperative systems has been studied in Yuksel and Erkip (2007).

More applied issues of wireless relay system have been discussed in Sendonaris et al. (2003) in terms of user cooperation diversity, where the key idea is to enable multiple terminals (source and/or relays) to cooperate with each other to create a virtual antenna array that provides some form of spatial diversity. The authors of Sendonaris et al. (2003) showed that user cooperation is beneficial in terms of increasing system throughput and cell coverage, as well as decreasing sensitivity to channel variations.

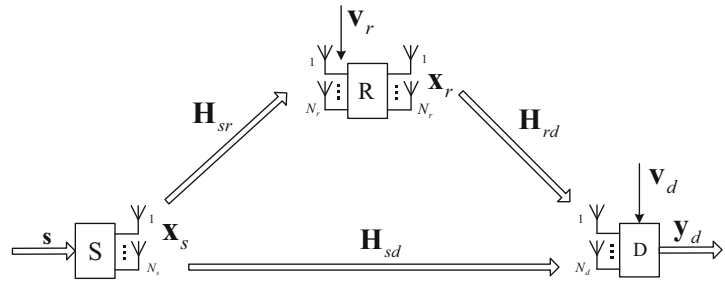
Recent surveys on topics in *amplify-and-forward* (AF) MIMO relay systems can be found in Sanguinetti et al. (2012). A testbed implementation of MIMO relay system was reported in Maltsev et al. (2010). A general framework of optimizing the source and relay precoding matrices has been developed in Rong et al. (2009). There are special issues in *IEEE Journal on Selected Areas in Communications* on the topic of MIMO relay (Hua et al. 2012).

Foundations

The simplest MIMO relay system consists of three nodes as illustrated in Fig. 1, where one

MIMO Relay, Fig. 1

Block diagram of a dual-hop MIMO relay system



source node (S) transmits information to one destination node (D) with the aid of a relay node (R). Each node has multiple transmit/receive dimensions. Let us denote the number of dimensions at node i as N_i , $i \in \{s, r, d\}$. Mathematically the channel response from one node to another is represented through a matrix. As an example, the source-relay, relay-destination, and source-destination channels are represented by matrices $\mathbf{H}_{sr}$, $\mathbf{H}_{rd}$, and $\mathbf{H}_{sd}$, respectively. MIMO relay channel is an important extension of single-hop MIMO channel, which has been extensively studied in the last decade. Similar to a single-hop MIMO channel, the multiple transceiving dimensions (degrees of freedom) in a MIMO relay channel can be used to achieve various trade-offs between diversity gain and multiplexing gain (Yuksel and Erkip 2007), where the diversity gain helps improving the system reliability, while the multiplexing gain is mainly used to increase the system throughput.

Since each hop of the relay system in Fig. 1 forms a MIMO channel, the input-output relationship from node $j \in \{s, r\}$ to node $i \in \{r, d\}$ is given by

$$\mathbf{y}_i = \mathbf{H}_{ji}\mathbf{x}_j + \mathbf{v}_i \quad (1)$$

where $\mathbf{x}_j$ is the transmitted signal vector at node j with a dimension of N_j and $\mathbf{y}_i$ and $\mathbf{v}_i$ are $N_i \times 1$ vectors of received signal and noise at node i , respectively.

Compared with single-hop MIMO systems, dual-hop or multihop MIMO systems provide more freedoms to system designers. In particular, for an optimal dual-hop relay system performance, both of the transmitted signal vectors $\mathbf{x}_s$ and $\mathbf{x}_r$ from two different nodes need to be optimized, and both of the received signal vectors $\mathbf{y}_r$

and $\mathbf{y}_d$ at two different nodes need to be optimally processed. On the other hand, such a flexibility also makes the optimization problems much more challenging than those for single-hop MIMO systems. In fact, most of the optimization problems in MIMO relay systems are highly nonconvex involving multiple matrix/vector variables. As a matter of fact, except for some upper bounds (Wang et al. 2005), the exact capacity of a MIMO relay channel is still not available.

A relay can be half-duplex or full-duplex. A half-duplex relay receives and transmits in two separate time/frequency channels. A full-duplex relay receives and transmits at the same time and same frequency (Cirik et al. 2014).

Relay Strategies

An important part of a relay system is known as relay strategy, which determines how the relay processes $\mathbf{y}_r$ to generate $\mathbf{x}_r$. Mathematically, the relay strategy can be represented by a function of $\mathbf{x}_r = \mathbf{f}(\mathbf{y}_r)$, where $\mathbf{f}$ represents the relay strategy. Interestingly, so far, there is no relay strategy that works best under all scenarios. Generally speaking, there are two main categories of relay strategies: *regenerative* relay and *non-regenerative* relay. In a regenerative strategy, the relay node first extracts out (decodes) the information from $\mathbf{y}_r$. Then, the relay node generates $\mathbf{x}_r$ by re-encoding the information. This strategy is also known as *decode-and-forward* (DF). A relaxed form of this strategy is called *compress-and-forward* where the signal received at the relay node is partially compressed before it is

forwarded. There are also other variations of this strategy (Kramer et al. 2005).

For a *non-regenerative* strategy, the relay node only amplifies (including a possible linear transformation) and retransmits its received signals, without attempting to decode the information-carrying symbols. Thus, a non-regenerative relay is also referred to as AF relay. The authors of Fan and Thompson (2007) compared the performance-complexity trade-offs of non-regenerative and other MIMO relay techniques. Both regenerative and non-regenerative strategies are affected by the noise at the relay node, although in different ways. The complexity and the processing delay of a non-regenerative strategy are generally much smaller than those of a regenerative strategy. The non-regenerative strategy is also believed by many to provide a better trade-off between benefits and implementation costs.

With a linear AF relay, we can in general write $\mathbf{x}_r = \mathbf{F}_r \mathbf{y}_r$, where $\mathbf{F}_r$ is the relay amplifying (precoding) matrix.

Key Applications

Relay nodes are needed in situations where the path loss between source and destination is too high, and/or the transmission power from the source is too limited by regulation and/or hardware constraints. Relay nodes are easy to install to extend the reach of a backbone network. Special relay nodes can be equipped with multiple antennas to compensate the limitation of most user terminals such as handsets. Cooperative relay nodes can form a virtual multi-antenna relay (Sendonaris et al. 2003), which increases the spatial diversity order of the source-relay-destination channel and thus improves the reliability of communication.

MIMO relay has applications in many physical forms of communication channels, such as:

- Multi-antenna multihop wireless networks, which are perhaps the most commonly referred to in the context of MIMO relays.

Multihop wireless backhaul networks are being considered in several industry standards such as IEEE802.16j (Genc et al. 2008).

- Underwater acoustic relay channel. Since the bandwidth of underwater acoustic channel is inversely proportional to the transmission distance, MIMO relay techniques can enhance the capacity of underwater acoustic communication systems over a long distance (Al-Dharab et al. 2013).
- Wireline DSL relay channel, which can be modeled as MIMO relay channel due to the crosstalk arising from the electromagnetic coupling between neighboring twisted-pairs.
- Powerline relay channel. Recently, cooperative relay communication technology has been adopted into the indoor powerline communication environment (Wu and Rong 2015).

Cross-References

- [Network MIMO](#)
- [Relaying in LTE-Advanced](#)

References

- Al-Dharab S, Uysal M, Duman TM (2013) Cooperative underwater acoustic communications. *IEEE Commun Mag* 51:146–153
- Cirik AC, Rong Y, Hua Y (2014) Achievable rates of full-duplex MIMO radios in fast fading channels with imperfect channel estimation. *IEEE Trans Signal Process* 62:3874–3886
- Cover TM, El Gamal AA (1979) Capacity theorems for the relay channel. *IEEE Trans Inf Theory* 25:572–584
- Fan Y, Thompson J (2007) MIMO configurations for relay channels: theory and practice. *IEEE Trans Wireless Commun* 6:1774–1786
- Genc V, Murphy S, Yu Y, Murphy J (2008) IEEE 802.16j relay-based wireless access networks: an overview. *IEEE Wireless Commun* 15:56–63
- Hua Y, Bliss DW, Gazor S, Rong Y, Sung Y (2012) Guest editorial: theories and methods for advanced wireless relays – issue I. *IEEE J Sel Areas Commun* 30:1361–1367
- Kramer G, Gastpar M, Gupta P (2005) Cooperative strategies and capacity theorems for relay networks. *IEEE Trans Inf Theory* 51:3037–3063
- Maltsev A, Khoryaev A, Lomayev A, Maslennikov R, Antonopoulos C, Avgeropoulos K, Alexiou A, Boccardi F, Hou Y, Leung KK (2010) MIMO and

multihop cross-layer design for wireless backhaul: a testbed implementation. *IEEE Commun Mag* 48:172–179

Rong Y, Tang X, Hua Y (2009) A unified framework for optimizing linear nonregenerative multicarrier MIMO relay communication systems. *IEEE Trans Signal Process* 57:4837–4851

Sanguinetti L, D'Amico AA, Rong Y (2012) A tutorial on the optimization of amplify-and-forward MIMO relay systems. *IEEE J Sel Areas Commun* 30:1331–1346

Sendonaris A, Erkip E, Aazhang B (2003) User cooperation diversity – Part I and Part II. *IEEE Trans Commun* 51:1927–1948

van der Meulen EC (1971) Three-terminal communication channels. *Adv Appl Prob* 3:120–154

Vanroose P, van der Meulen EC (1992) Uniquely decodable codes for deterministic relay channels. *IEEE Trans Inf Theory* 38:1203–1212

Wang B, Zhang J, Høst-Madsen A (2005) On the capacity of MIMO relay channels. *IEEE Trans Inf Theory* 51:29–43

Wu X, Rong Y (2015) Joint terminals and relay optimization for two-way power line information exchange systems with QoS constraints. *EURASIP J Adv Signal Process* v2015:84

Yuksel M, Erkip E (2007) Multi-antenna cooperative wireless systems: a diversity multiplexing tradeoff perspective. *IEEE Trans Inf Theory* 53:3371–3393

MIMO Relay Channel

► [MIMO Relay](#)

MIMO Relay Network

► [MIMO Relay](#)

MIMO Satellite

► [Satellite MIMO](#)

MIMO Signal Demodulation

► [MIMO Detection](#)

MIMO Signal Equalization

► [MIMO Detection](#)

Mining Task Offloading in Wireless Blockchain Networks

Mengting Liu¹, Richard Fei Yu^{2,3}, Yinglei Teng¹, Victor C. M. Leung⁴, and Mei Song¹

¹School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

²Carleton University, Ottawa, Canada

³Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

⁴The University of British Columbia, Vancouver, BC, Canada

Synonyms

[Computation offloading](#); [Wireless blockchain networks](#)

Definitions

Mining task offloading is a promising solution to address the computation-intensive mining tasks by offloading them to edge computing nodes using mobile edge computing technology.

Historical Background

Blockchain, the technology underpinning cryptocurrency, has been widely applied to many fields, such as security, smart cities, and supply chain UK Government (2016) and Iansiti and Lakhani (2017). However, the application of blockchain technology into wireless mobile networks is hindered by a main challenge brought by a computational process called

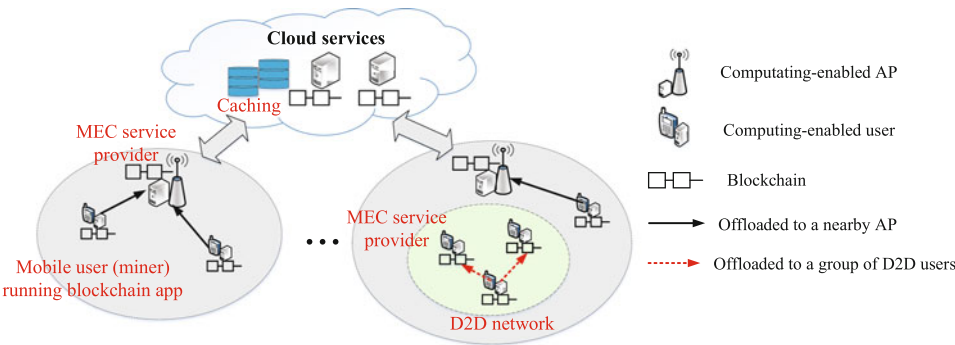
mining. Specifically, in order to confirm and secure the integrity and validity of transactions, the participants (a.k.a. miners) need to solve a computational difficult problem, i.e., the proof-of-work puzzle Beccuti (2017) and Kiayias et al. (2016), which sets an extremely high demand for the computational capabilities and storage availability of the miners.

To release the burden of the miners, *Mobile Edge Computing* (MEC) is considered as a promising solution that brings network resources closer to the end users, which can effectively reduce the energy consumption and shorten the transmission delay Stanciu (2013) and Xiong et al. (2017). As a key enabler of MEC, nearby resource-rich access points (APs) can process the users’ requests as a substitute of clouds, which is also called cloudlet and able to reduce the transmission delay due to its proximity to users Mao et al. (2017b), Wang et al. (2016), and Yu et al. (2016). Nonetheless, due to its limited computation capacity and storage space, it is difficult for each AP to handle a large amount of requests from the users at a time. Moreover, AP computation offloading may lose its advantages when the wireless channel is in deep fading or the user is far away from the AP. To avoid the global network bottlenecks, device-to-device (D2D) computation offloading acts as an additional type of edge computing is playing an increasing important role in enhancing MEC performance by exploiting surplus computing resources in neighboring user devices Mao et al. (2017a).

Therefore, there are mainly two offloading modes. One is “offloaded to the nearby AP” where the powerful computing capability enabled by the APs can be leveraged to enhance the task offloading performance. The other is “offloaded to a group of nearby users” (a.k.a. multiuser cooperative edge computing) where a group of users can share their computational resources to aid the computation process through D2D computation offloading. Accordingly, in MEC-enabled wireless blockchain networks, the mobile miners can resort to the nearby edge nodes (APs or mobile users) to perform the computation-intensive tasks.

Foundations

Since blockchain in general are currently employed in wired networks, a framework of wireless blockchain network with MEC is necessary to perform the mining task offloading. Figure 1 gives a brief framework of MEC-enabled wireless blockchain networks Liu et al. (2018), where both the APs and mobile users have a computation capacity to provide task-execution services. Accordingly, the computational difficult mining task can be offloaded in two ways: (1) The miners can offload the full mining task to a nearby AP. (2) The miners can divide the whole mining task into several parts and separately forward them to a group of nearby mobile users through D2D links.



Mining Task Offloading in Wireless Blockchain Networks, Fig. 1 An illustration of MEC-enabled wireless blockchain networks

Applying the blockchain technology to wireless networks induces several issues related to the resource allocation and management, which is an ever-popular topic in wireless networks. The most representative ones include:

- Offloading scheduling Liu et al. (2018): For the miners in wireless blockchain networks, offloading scheduling scheme including the offloading mode selection and offloading task allocation among two offloading modes is crucial to guarantee the miners' demands as well as balance the computation load of edge computing nodes.
- Resource pricing Xiong et al. (2017): In blockchain-based networks, the price setting of resources, an indispensable part in the incentive mechanism, plays an important role in coordinating the wireless resource allocation.
- Content caching: For the miners, some information related with the block (like the cryptographic hashes of blocks) and computation results need to be stored for future request. The caching strategy of miners (where to cache and which part to cache) coupled with the storage space allocation is a key factor for wireless blockchain networks.

Key Applications

Mining task offloading provides an effective way to address the computational intensive mining tasks for miners, which can facilitate the application of blockchain technology into wireless networks.

Cross-References

- [Integrated System of Networking, Caching, and Computing](#)
- [Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks](#)
- [Mobile Edge Computing: Low Latency and High Reliability](#)

References

- Beccuti J (2017) The bitcoin mining game: on the optimality of honesty in proof-of-work consensus mechanism. Tech. rep
- Iansiti M, Lakhani KR (2017) The truth about blockchain the truth about blockchain. Tech. rep
- Kiayias A, Koutsoupias E, Kyropoulou M, Tselekounis Y (2016) Blockchain mining games. In: Proceedings of the 2016 ACM conference on economics and computation, Maastricht, The Netherlands
- Liu M, Yu FR, Teng Y, Leung VCM, Song M (2018) Joint computation offloading and content caching for wireless blockchain networks. In: Infocom'18 workshops, IEEE publication, Honolulu, HI
- Mao Y, You C, Zhang J, Huang K, Letaief KB (2017a) A Survey on mobile edge computing: the communication perspective. IEEE Commun Surv Tutorials 19(4):2322–2358
- Mao Y, Zhang J, SS H, LK B (2017b) Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems. IEEE Trans Wirel Commun 16(9):5994–5600
- Stanciu A (2013) Blockchain based distributed control system for edge computing. In: 21st international conference on Control Systems and Computer Science (CSCS), IEEE publication, Bucharest, Romania
- UK Government (2016) Distributed ledger technology beyond block chain. Tech. rep
- Wang Y, Sheng M, Wang X, Wang L, Li J (2016) Mobile-edge computing: partial computation offloading using dynamic voltage scaling. IEEE Trans Commun 64(10):4268–4282
- Xiong ZH, Zhang Y, Niyato D, Wang P, Han Z (2017) When mobile blockchain meets edge computing: challenges and applications. Comput Sci
- Yu Y, Zhang J, Letaief KB (2016) Joint subcarrier and cpu time allocation for mobile edge computing. In: IEEE global communications conference, IEEE publication, Washington, DC

MIP

- [Mobile IP](#)

MIPv4

- [Mobile IP](#)

MIPv6

► Mobile IP

Mixed Network

Sanaa Taha¹ and Xuemin (Sherman) Shen²

¹Information Technology Department, Faculty of Computers and Information, Cairo university, Cairo, Egypt

²Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

Synonyms

Digital pseudonyms; Mixnet; Return address; Unreachable electronic mail

Definitions

Mixed network, is a message routing protocol supporting sender and receiver anonymity by using a group of proxy servers, called mixes or stages. At each mix, a batch of inputs is mixing together and reordering to exit in a different order.

Introduction

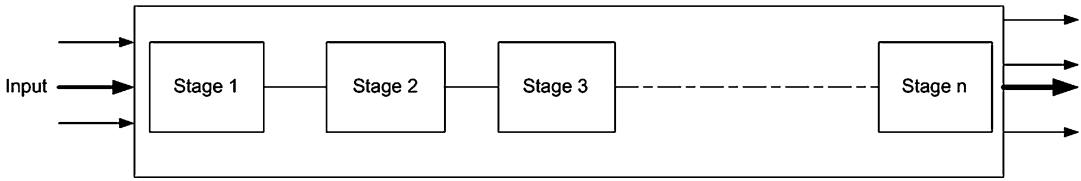
Mixed network, or mixnet, is a message routing protocol supporting sender and receiver anonymity by using a group of proxy servers, called mixes or stages. At each mix, a batch of inputs is mixing together and reordering to exit in a different order. The first mixnet protocol was proposed by David Chaum in 1981 to achieve anonymity in electronic voting (e-voting) application. Consequently, many anonymity networks have been proposed, such as Crowds (Reiter and Rubin 1998), MixMaster (Cornelius et al. 2008), Freenet (Clarke et al. 2000), Tarzan (Freedman and Morris 2002), Tor (Dingeldine et al. 2004), GUNet (<https://gnunet.org/>), and I2P (<https://geti2p.net/>). Additionally, the mixnet also provides message integrity and verifiability. From performance perspective, the mixnet should acquire less network latency and high throughput. The goal of mixnet is either to support one (sender) or two-way (sender and receiver) anonymity.

In the context of wireless networks, many types of applications use mixnet routing, including e-voting, anonymous e-mails, and location privacy in wireless networks (Lu et al. 2009). For secure and error-free election, the e-voting system requires verification of voter eligibility, ballot integrity, and tally accuracy. For anonymous e-mails, mixnet should apply reliability to anonymous Internet communications to assure detecting any misuse of anonymous emails senders. Also, location privacy applications can be efficiently used with anonymous sender, such as Radio frequency identification (RFID) tags anonymity (Lin et al. 2010).

Applications employing mixnet routing protocol face passive and active attacks (Duddu and Samanta 2018). Passive attacks, also called traffic analysis, have the goal of correlating inputs to corresponding outputs at each mix stage, while active attacks, also called traffic manipulation, have the goal of controlling one or more mix stages to attack message integrity.

Mixnet Types

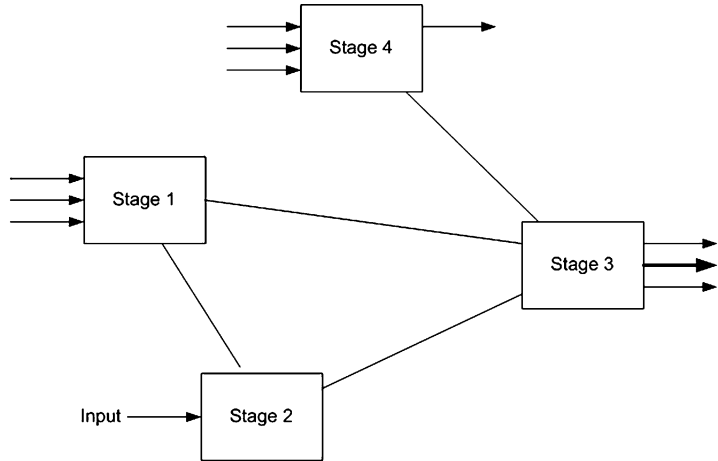
The topology of the mixnet can be cascade or free-route (Taha and Shen 2013), as illustrated in Figs. 1 and 2, respectively. On one hand, the mixes in the cascade mixnet are ordered in sequential and every batch of inputs should go through all mixes in the network. On the other hand, the free-route mixnet is not restricted to a fixed group of mixes and all inputs may traverse different paths in the network. Depending on mixing operation types, the mixnet implements one (or the two) of those topology types. Additionally, those mixing operations are categorized into decryption, hybrid, re-encryption,



Mixed Network, Fig. 1 Cascade mixnet topology

Mixed Network, Fig. 2

Free route mixnet topology



and universal re-encryption mixnet (Sampigethaya and Poovendran 2006).

In decryption mixnet, the sender uses each mix's public key to repeatedly encrypt the message, starting with the first mix to the receiver and ending with the first mix to the sender. In each encryption, the sender adds the next mix's address; hence, each mix knows the next mix's address only when decrypting the message. Only the last mix knows the receiver, x , of the message when decrypting the message. Equation (1) shows the encrypted message transferred from the sender, s .

$$E_K(m, r) = A_1 \| E_{K_1} (A_2 \| E_{K_2} (A_3 \| \dots \| E_{K_{n-1}} (A_n \| E_{K_n} (A_x \| E_{K_x} (m) \| r_n) \| r_{n-1}) \dots \| r_2) \| r_1) \quad (1)$$

where $E_K(m, r)$ is an encryption operation, El Gamal or RSA encryption, using a public key K , A_i is the address of a mix i , and r_i is a random number generated by a mix i . This type of mixing is not only used for sender anonymity,

but it also can be used for two-way anonymous communications to achieve receiver anonymity. This can be achieved by adding to the previous equation the Return Path Information (RPI), which includes the return information encrypted using sender's public key, K_s . However, using many public key operations may limit the mixnet high-latency applications.

For hybrid mixing, the public key is used only to encrypt the shared keys, sk , between the sender and every mix in the network. The shared keys are used in encrypting the transmitted messages as illustrated in Eq. (2)

$$E_K(m, r, sk) = A_1 \| E_{K_1} (sk_1) \| E_{sk_1} [A_2 \| E_{K_2} (sk_2) \dots \| E_{sk_n} [A_n \| E_{K_n} (A_x \| E_{K_x} (m) \| r)] \dots] \quad (2)$$

For re-encryption mixnet type, the encrypted message does not have to go through all mixes, i.e., free-route mixnet is more suitable for this type. Additionally, and unlike other mixnet operation types, the length of the message does not

changed as the message goes through mixes. The reason of this fixed length message is that the sender uses an encryption key, which is a combination of all mixes' keys. In other words, Eq. (3) shows the encryption message produced at the sender

$$E_K(m, r) = (g^r \| (A_x \| m) K^r) \quad (3)$$

where g is the generator of a finite group Z_p , and K is the public key of the mixnet, calculated as follows:

$$K = \prod_{j=1}^n K_j = \prod_{j=1}^n g^{d_j} = g^{\sum_{j=1}^n d_j} \quad (4)$$

However, to achieve multicasting transmission, (i.e., transmitting to multiple receivers), all mixes in the re-encryption mixnet must share their key and perform decryption after mixing to get the address of the receiver. Such dependency can be avoided when using the universal re-encryption mixnet, where the sender transmits two encrypted messages: one for the original message and the other for the receiver's public key used to encrypt this original message, as illustrated in Eq. (5).

$$E_K(m, r) \| E_K(1, r') = (g^r \| m K^r) \| (g^{r'} \| K^{r'}) \quad (5)$$

Table 1 shows a comparison between the different mixing types of mixnet

Degree of Anonymity

The degree of anonymity, d , is a measure of anonymity level achieved when applying the mixnet protocol, and this level varies depending on the supported applications. For instance, in e-voting, the voter's identity should be totally hidden, while in electronic payment system, seller's identity can be revealed by an authorized party in order to solve any dispute.

According to Reiter and Rubin (1998), the degree of anonymity is defined as: $d = 1 - p$,

where p is an attacker probability about a potential sender. Additionally, Entropy is employed as a measure of the degree of anonymity as follows (Diaz 2006):

$$H(x) = \sum_{i=1}^N \left[p_i \log \frac{1}{p_i} \right],$$

where N is number of nodes inside the network and p_i is the probability associated with a node i . For a uniform distribution, all nodes have an equally likely probability of $\frac{1}{N}$ and Entropy records a maximum, H_M . Therefore, the degree of anonymity can be calculated as follows: $= \frac{H(x)}{H_M}$.

The effective anonymity set size, N , is another factor in measuring the degree of anonymity, which increases as the set size increases.

Mixnet-Based Wireless Applications

Shuffling protocol (Peng 2011) is employed to implement the mixnet, in which re-encryption followed by shuffling proof operations is applied on the group of ciphertext. Shuffling proof is a complex process, used to prove that the permutation of the input results the correct output. Shuffling protocol should support correctness, soundness, and zero knowledge privacy. Correctness means that shuffling inputs specific batches to get intended output, soundness means to verify the correctness of shuffling, and zero knowledge privacy means that it is impossible to reveal the permutation used in shuffling operation.

Another mixnet-based project is the Participatory sensing system (PSS) (Peng 2011), which is a data collection, processing, and dissemination system that is cost-effective and reliable. The system monitors a set of assets, called point of interest (POIs), by a group of entities, such as mobile nodes, which are participating to observe the assets and send their observed information to the application server. Based on the idea of mixnet, if the observer senses the attribute of POI, it sends an anonymous observation report to the application server. In order to anonymously send the report, the observer entity transmits the report

Mixed Network, Table 1 Mixnet types comparison

	Decryption	Hybrid	Re-encryption	Universal re-encryption
Mix type	Cascade	Cascade	Cascade and free route	Free route
Multiple receiver	Not applied	Not applied	Could be applied	Could be applied
Anonymity	One-way, could be two-way by adding RPI	One-way, could be two-way by adding RPI	Two-way is possible	One-way only
Message Integrity	Using public-key encryption	Using public and symmetric keys	Using El Gamal encryption	Using El Gamal encryption
Verifiability (Puiggali 2010; Chen 2007)	Using public key verification techniques	Using public key verification techniques	Using El Gamal verifications	Using El Gamal verifications
Fault Tolerance	Not applied	Not applied	applied	applied
Latency	High	Higher	Medium	Low
Throughput	lower	Low	Medium	High
Scalability	Not applied	Not applied	Could be enhanced	Could be enhanced
Efficiency	Low	Low	Medium	Medium

to another entity in the network and identifies specific number of hops for the message to circulate before receiving to the application server.

The Onion Routing (Tor) (Burke et al. 2006) is another predecessor for mixnet to achieve anonymous communications, in which the message has layers of encryption, and one layer is peeled at a time as the message traverses through Tor volunteering servers. However, Tor faces obfuscations of limited capacity and performance due to relaying on volunteering servers. In (The Tor project 2003), a Software-Defined Networking (SDN)-based Onion Routing (SOR) is proposed as a novel Cyber anonymity using Tor (Elgzil et al. 2017). The goal is to leverage the large capacities, strong connectivity, and economies of scale inherent to commercial data centers, by building onion routing tunnels over anonymity service providers.

Cross-References

► Proxy Mobile IPv6

References

Burke J, Estrin D, Hansen M, Parker A, Ramanathan N, Reddy S, Srivastava MB (2006) Participatory sensing.

- In: Proceedings of ACM conference on embedded networked sensor systems, Boulder, Colorado, USA
- Chaum D (1981) Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun ACM* 4(2):84–88
- Chen J (2007) Verifiable mixnets techniques and prototype implementation. Master's thesis, Darmstadt University of Technology
- Clarke I, Sanberg O, Wiley B, Hong TW (2000) Freenet: a distributed anonymous information storage and retrieval systems. In: Proceeding of the ICSI workshop on design issues in anonymity and unobservability, Springer
- Cornelius C, Kapadia A, Kotz D, Peebles D, Shin M, Triandopoulos N (2008) Anonymsense: privacy-aware people-centric sensing. In: Proceeding of the 6th international conference on mobile systems, applications, and services, Breckenridge, Colorado, pp 211–224
- Diaz C (2006) Anonymity metrics revisited. In: Dagstuhl seminar proceedings, IBFI, Schloss Dagstuhl, Germany. <http://drops.dagstuhl.de/opus/volltexte/2006/483>
- Dingeldine R, Mathewson N, Syverson P (2004) Tor: the second-generation onion router. Naval Research Lab, Washington, DC
- Duddu V, Samanta D (2018) Network and security analysis of anonymous communication networks. *arXiv preprint arXiv:1803.11377*. Available online, <https://arxiv.org/abs/1803.11377>
- Elgzil A, Chow CE, Aljaedi A, Alamri N (2017) Cyber Anonymity based on software-defined networking and onion routing. In: Proceeding of IEEE conference on dependable and secure computing, Taipei, pp 358–365
- Freedman MJ, Morris R (2002) Tarzan: a peer-to-peer anonymizing network layer. In: Proceeding of the 9th ACM conference on computer and communications security, Washington, DC, USA

GNUnet. Available online, <https://gnunet.org/>

I2P. Available online, <https://geti2p.net/>

Lin X, Lu R, Kwana D, Shen X (2010) REACT: an RFID-based privacy-preserving children tracking scheme for large amusement parks. *Comput Networks* (Elsevier) 54(15):2744–2755

Lu R, Lin X, Zhu H, Ho PH, Shen X (2009) A novel anonymous mutual authentication protocol with provable link-layer location privacy. *IEEE Trans Veh Technol* 58(3):1454–1466

Peng K (2011) Survey, analysis and re-evaluation – how efficient and secure a mix network can be. In: *Proceeding of IEEE 11th international conference on Computer and Information Technology (CIT)*, Paphos, Cypru, pp 249–254

Puiggali J (2010) Universally verifiable efficient re-encryption mixnet. *Krimmer and Grimm* [32], pp 241–254

Reiter M, Rubin A (1998) Crowds: anonymity for web transactions. *ACM Trans Inf Syst Secur* 1(1):66–92

Sampigethaya K, Poovendran R (2006) A survey on mix networks and their secure applications. In: *Proceeding of the IEEE*, vol 94, No. 12, Dec 2006

Taha S, Shen X (2013) ALPP: anonymous and location privacy preserving scheme for mobile IPv6 heterogeneous networks. *Secur Commun Networks* 6(4):401–419

The Tor project (2003.) <https://www.torproject.org/>. Accessed Nov 2015

Mixnet

- [Mixed Network](#)

mmWave Channel Campaigns

- [Millimeter Wave Channel Measure](#)

mmWave Channel Measurement

- [Millimeter Wave Channel Measure](#)

mmWave Channel Sounding

- [Millimeter Wave Channel Measure](#)

Mobile Access Gateway

- [Proxy Mobile IPv6](#)

Mobile Big Data

- [Data-Driven Mobile Social Networks](#)

Mobile Caching

- [Wireless Edge Caching](#)

Mobile Cloud Computing

- [Mobile Edge Computing: Low Latency and High Reliability](#)

Mobile Content Delivery

- [Wireless Edge Caching](#)

Mobile Content Delivery Architecture

- [Mobile Content Distribution Architecture](#)

Mobile Content Distribution Architecture

Junfeng Xie

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Synonyms

[Content delivery architecture in mobile cellular network](#); [Content distribution architecture in](#)

mobile cellular network; Mobile content delivery architecture

Definition

Mobile content distribution architecture is a promising technique to optimize the mobile content delivery and decrease duplicate transmission by offering flexible storage capability in mobile network, which can utilize network resources in a more efficient manner and improve end users' Quality of Experience (QoE).

Foundations

In recent years, to reduce mobile core network traffic and deliver content efficiently over the cellular network, many content distribution architectures have been proposed. Mobile content delivery network (CDN), mobile edge computing (MEC), and information-centric networking (ICN) are the representative architectures. So in this section, we briefly present the research on the three emerging content distribution techniques.

Mobile CDN

In the past decade, CDN has become the most popular and widely used content distribution architecture in the current wired Internet. Today, a large fraction of Internet traffic is delivered by CDN. Although CDN works very well in wired networks, it cannot be deployed in mobile networks directly. CDN is not suitable for mobile networks for two reasons. Firstly, CDN systems just help clients to select the "best" server at the initial stage and do not consider clients' mobility. Secondly, DNS-based schemes are used in traditional CDN systems to select cache servers and perform load balancing. Nevertheless, in a highly volatile mobile environment, user equipment (UE)'s location and IP address can change over time rapidly, which makes the address not always truly reflect the position of a UE. Thus, how to deploy CDN in mobile network becomes an interesting research direction.

One feasible mobile CDN implementation solution to address the scalability and mobility issues of the mobile networks is jointing distributed mobility management (DMM) and CDN (Liebsch and Yousaf 2013; Munaretto et al. 2014). DMM has been proposed by the Internet Engineering Task Force (IETF) as a new paradigm for flat networks in 5G mobile architectures to enable IP address continuity in DMM. Other mobile CDN implementation solutions have also been studied. Literatures (Dramitinos et al. 2013; Pentikousis et al. 2013) utilize the emerging software defined networking (SDN) approach to increase the operators' innovation potential and support video services more effectively.

MEC

With the evolution of mobile base stations and the emergence of cloud computing technology, edge computing as a novel paradigm has attracted widespread attention. In recent years, a few edge computing architectures have been presented, such as Cloudlet, Edge Computing, Fog Computing, Mobile Cloud Computing (MCC), Mist Computing, and MEC (Mach and Becvar 2017). Among all these edge computing architectures, MEC is the most popular architecture. MEC has been proposed by the standards institute ETSI (European Telecommunications Standards Institute) as a promising technology to reduce the load of mobile core network by shifting computational capability from the core network to the mobile edge (i.e., the radio access network). The key characteristics of MEC are proximity, lower latency, and location awareness. The MEC white paper has been issued in 2014 (Patel et al. 2014). MEC is an emerging technology, which is very likely to be applied in 5G mobile networks. Thus, there is a trend of integrating content distribution with MEC to improve content delivery efficiency and reduce content delivery delay.

ICN

Over the past several years, the research on the content distribution architecture is from two approaches: "clean slate" and "evolutionary".

Mobile CDN and MEC all belong to the “evolutionary” approaches. However, ICN (Xylomenos et al. 2014; Xie et al. 2017c) is the representative of “clean slate” approaches. The key concept of ICN is the content-centric communication model instead of the end-to-end communication model in traditional networks. So far a few ICN architectures have been proposed, such as Data-Oriented Network Architecture (DONA), Publish Subscribe Internet Technology (PURSUIT), Architecture and Design for the Future Internet (4WARD), Scalable and Adaptive Internet Solutions (SAIL), and Named-Data Networking (NDN). Although the concept of ICN has already attracted great attention in academia and industry, the deployment of ICN needs to update all network elements, user devices, as well as origin servers to be ICN-aware. In this case, the pure commercial ICN deployment would not be done in the near future. So literatures (Veltri et al. 2012; Vahlenkamp et al. 2013; Chanda and Westphal 2013; Chang et al. 2014) have studied the issue on how to deploy the ICN architectures in the legacy IP network.

Future Directions

Despite the potential vision of these mobile content distribution architectures, many significant research challenges remain to be addressed by future research efforts. In this section, we discuss some of these challenges and present some future research directions.

Adaptive Video Content Distribution over Cellular Networks

In mobile cellular networks, due to the varying wireless network conditions and the diversity of end-user devices, different users should be served with different data rates to enhance the users’ QoE. Researchers have proposed various techniques to provide better QoE for mobile users by adapting multimedia content over wireless channels real-timely. Recently, the Scalable Video Coding (SVC) (Xie et al. 2017b) and HTTP Adaptive Bit Rate (ABR) streaming (Xie et al. 2017a) have been considered as the main

solutions. In SVC, each video is encoded into one mandatory base layer and several optional enhancement layers. The enhancement layers build upon the base layer and provide multiple qualities. In HTTP ABR streaming, a video is divided into a series of small video chunks with a typical time duration of 2–10 s, and each video chunk is encoded into multiple bitrates (i.e., qualities). A client can adaptively request a video content with different qualities according to its wireless network condition. Caching adaptive video content in Radio Access Network (RAN) is a promising way to provide better content delivery service and to improve content distribution efficiency (Xie et al. 2016). In this scenario, how to allocate limited cache resource for adaptive video contents has become an important issue. Therefore, it is necessary to design effective cache resource allocation schemes to optimize the adaptive video content placement.

HTTPS Traffic Processing Strategies

Most of the current works on content distribution solutions are studied especially for HTTP traffic. With increased emphasis on privacy, security, and regulatory compliance, organizations and companies are more and more dependent on encryption technologies to protect information, among which TLS/SSL have become the standard of choice for providing secure applications. For instance, TLS/SSL are used by HTTPS to ensure secured transactions. On the other hand, with the rapid development of processing technologies, the cost to encrypt data is falling rapidly. Unfortunately, once an encrypted connection between a client and a server is established, it is difficult to manage, accelerate, or audit activity due to the private nature of the TLS/SSL tunnel. Therefore, it is necessary to study how to solve the HTTPS traffic processing problem in content distribution framework in the future.

Combining with Network Function Virtualization and Cloud Computing

With the rapid development of processing and virtualization technologies, computing resources

have become cheaper, more powerful, and more ubiquitously available than ever before. Network Function Virtualization (NFV) is a promising technology which decouples network functions from the underlying specialized hardware to make the network architecture more flexible and open. The NFV and cloud computing technologies draw ISPs' attention because they can reduce operators' capital expenditures for scaling up the network and make network reconfiguration quick and adaptive. Therefore, the deployment of content distribution architecture combining with NFV and cloud computing technologies is another future research direction.

Key Applications

Mobile content distribution architectures can be deployed in mobile networks to provide better content delivery service and improve end users' QoE.

Cross-References

- [Enhancing QoE-aware Wireless Edge Caching with Software-defined Wireless Networks](#)
- [Integrated System of Networking, Caching, and Computing](#)
- [Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks](#)
- [Quality of Experience for Wireless Video Streaming](#)
- [Wireless Video Delivery](#)
- [Wireless Video Streaming](#)

References

- Chanda A, Westphal C (2013) ContentFlow: mapping content to flows in software defined networks. CoRR abs/1302.1493
- Chang D, Kwak M, Choi N, Kwon T, Choi Y (2014) C-flow: an efficient content delivery framework with OpenFlow. In: Proceedings of the IEEE ICOIN'14, Phuket
- Dramitinos M, Zhang N, Kantor M, Costa-Requena J, Papafili I (2013) Video delivery over next generation cellular networks. In: Proceedings of the IEEE CNSM'13, Zurich
- Liebsch M, Yousaf FZ (2013) Runtime relocation of CDN serving points enabler for low costs mobile content delivery. In: Proceedings of the IEEE WCNC'13, Shanghai
- Mach P, Becvar Z (2017) Mobile edge computing: a survey on architecture and computation offloading. IEEE Commun Surv Tutor 19(3):1628–1656
- Munaretto D, Giust F, Kunzmann G, Zorzi M (2014) Performance analysis of dynamic adaptive video streaming over mobile content delivery networks. In: Proceedings of the IEEE ICC'14, Sydney
- Patel M, Naughton B, Chan C, Sprecher N, Abeta S, Neal A (2014) Mobile-edge computing introductory technical white paper. White paper, Mobile-edge Computing (MEC) industry initiative
- Pentikousis K, Wang Y, Hu W (2013) Mobileflow: toward software-defined mobile networks. IEEE Commun Mag 51(7):44–53
- Vahlenkamp M, Schneider F, Kutscher D, Seedorf J (2013) Enabling information centric networking in IP networks using SDN. In: Proceedings of the IEEE SDN4FNS'13, Trento
- Veltri L, Morabito G, Salsano S, Blefari-Melazzi N, Detti A (2012) Supporting information-centric functionality in software defined networks. In: Proceedings of the IEEE ICC'12, Ottawa
- Xie J, Xie R, Huang T, Liu J, Yu FR, Liu Y (2016) Caching resource sharing in radio access networks: a game theoretic approach. Front Inf Technol Electron Eng 17(12):1253–1265
- Xie J, Xie R, Huang T, Liu J, Liu Y (2017a) Energy-efficient cache resource allocation and QoE optimization for HTTP adaptive bit rate streaming over cellular networks. In: Proceedings of the IEEE ICC'17, Paris, France, pp 1–6
- Xie J, Xie R, Huang T, Liu J, Liu Y (2017b) Energy-efficient content placement for layered video content delivery over cellular networks. In: Proceedings of the IEEE GLOBECOM'17, Singapore, pp 1–6
- Xie J, Xie R, Huang T, Liu J, Liu Y (2017c) ICICD: an efficient content distribution architecture in mobile cellular network. IEEE Access 5:3205–3215
- Xylomenos G, Ververidis C, Siris V, Fotiou N, Tsilopoulos C, Vasilakos X, Katsaros K, Polyzos G (2014) A survey of information-centric networking research. IEEE Commun Surv Tutor 16(2):1024–1049

Mobile Crowdsensing

- [Private Truth Discovery in Cloud-Empowered Crowdsensing Systems](#)

Mobile Crowdsensing: Issues and Challenges

Yang Liu

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Synonyms

Crowdsensing; Crowd sensing

Definitions

Mobile crowdsensing can be referred to as crowd sensing or crowdsensing, which is from the idea of crowdsourcing. Crowdsourcing comes from “wired” magazine which describes a novel distributed problem-solving and working model in which companies allocate tasks through the Internet, find ideas, or solve technological problems (Ganti et al. 2011). Recently, the idea of crowdsourcing has been combined with mobile sensing and uses the mobile devices as basic sensing units. The network forms a crowdsensing network, which realizes the distribution of sensing tasks and the collection of data.

Historical Background

At present, the Internet of Things has developed deeply, and the need for physical environment sensing is very thorough. With the development of wireless communication, smartphones, tablets, wearable devices, and mobile terminals such as in-vehicle sensing devices integrate more and more sensors. With the increase number of wireless mobile devices, more than a dozen people have been developing how to use specific consciously deployed sensors to provide sensing services. Also, the Internet of Things will provide more large-scale, complex, and comprehensive sensing services, thus entering a new era of development.

Foundations

In traditional networks, people usually consume sensing data. In crowdsensing, however, people serve both as a “consumer” of sensing data and as a “producer” of sensing data, using a popular new word, which is called “prosumer.” This basic human-centered feature brings unprecedented opportunities for IoT sensing and transmission. The specific performance is as follows:

- The costs of network deployment are lower. First, mobile devices do not require special deployment. Second, people help data sensing and transmission. As users who carry the devices arrive at different places, they can sense anytime and anywhere; on the other hand, due to the contacts between mobile users, they can use “storage-carry-forward” opportunistic transmission mode in an intermittently connected environment (Liu et al. 2018).
- Network maintenance is simpler. First, the users usually have better energy supply, computing, storage, and communication capabilities. Second, the mobile devices are usually kept by their holders, so they are usually in a better working state. For example, people can always charge mobile devices when their mobile phones are needed.
- The crowdsensing system is large scale. We need to recruit more users to participate in the system when there are more tasks.

Due to the above advantages, the crowdsensing network has become a new and important way to sense for the Internet of Things.

A typical crowdsensing is usually composed of a sensing platform and multiple users. Multiple servers are in the platform; the users can utilize various sensors embedded in the smart phones, e.g., GPS, accelerometer, gravity sensor, gyroscope, electronic compass, light distance sensor, microphone, cameras, etc., to collect various data and move through cellular networks and report the data. The workflow is as follows:

- **Task publishing:** data demanders publish tasks on the platform. There are two methods to assign tasks: pull-based method and push-based method. Pull-based method refers to that participants browse and retrieve tasks from the platform and actively register as workers of a task; push-based method refers to that the platform finds suitable candidates from all participants. No matter which task assignment method is adopted, when the participants accept the task, they will become a worker of the task;
- **Task execution:** one worker arrives at the designated place according to the task requirements and uses a specific application which automatically saves sensor data according to the task requirements. For real-time sensing tasks, workers must upload data immediately; for nonreal-time sensing tasks, workers are allowed to choose the most economical communication mode to upload data before the task deadline;
- **Data aggregation:** affected by the distributed sensing mode, there are a lot of redundant, i.e., low quality or repetitive data. In the data aggregation stage, the platform filters and optimizes the data according to the task requirements and selects the high-quality data set that meets the task requirements;
- **Result handover:** data demanders can download data aggregation results from the platform before or after the end of the task. Most mobile crowdsensing tasks are difficult to define data collection constraints accurately at the very beginning. Therefore, the platform allows data demanders to see data aggregation results before the end of the task, so that the data demander has the opportunity to adjust the task requirements before the end of the task.

Key Applications

Mobile crowdsensing can be applied in many important fields, such as intelligent transportation, public transportation, tourism encyclopedia, indoor positioning, information dissemina-

tion, news reporting, and event sensing. Some typical applications are listed below:

- **Intelligent transportation.** Gazetiki (Popescu et al. 2008) is an application that uses Wikipedia and web search to build a geographic encyclopedia. Gazetiki first identifies all geographical names and their coordinates through Wikipedia and classifies and sorts them. When users search geographic information, Gazetiki shows them detailed geographic information and relevant photos. ParkNet (Mathur et al. 2009) uses GPS and the ultrasonic sensor installed on the right door to detect the empty parking space and share the detection results; the literature in (Zhou et al. 2012) designs the bus arrival time prediction system under Android platform.
- **Shopping.** Mobishop (Sehgal et al. 2008) collects people's shop ticket photos, extracts commodity prices through OCR character recognition technology, and pushes the collected commodity price information to different customers, so that people can share commodity price information.
- **Entomology.** Lostadybug (Losey et al. 2012) is a project to collect Ladybug photos, which is used to study Ladybug species, living conditions, and habits. As of March 2017, the website has collected 38,000 data from all over the world.
- **Environmental monitoring.** Creekwatch (Kim et al. 2011) is developed by IBM to monitor river water quality. When people pass a stream or river, the creekwatch application will submit the status of the river in its current location, the status includes the amount of water and waste, and it will also submit some photos. Creekwatch shows people the status of rivers around the world through a website.
- **Emergency aid.** Wreckwatch (White et al. 2011) can detect the accident through the sensor of mobile phone in the car and then find passersby to take photos of the scene of the accident and send them to the rescue center, so that the rescue workers can judge the location of the accident and the emergency degree of the rescue timely and accurately.

- News reporting. When there is a special event, Mediascope (Jiang et al. 2013) will retrieve all kinds of photos from the mobile photo album shared by the witnesses on the scene for the purpose of news reporting. The retrieval conditions include image similarity and geographic information.
- Information dissemination. Fliermeet (Guo et al. 2014) collects photos of public posters in the city; calculates the relationship among people, posters, and places; constructs the preferences of different individuals for different types of posters; pushes the posters they like or need to different people; and improves the communication efficiency of urban information.
- Enjoying flowers. SakuraSensor (Morishita et al. 2015) judges the sakura blossom situation through the video clips people take in the car during the sakura blossom season and calculates the optimal route to visit the cherry blossom beauty according to people's driving path.
- Indoor locating and navigating. An image-aided positioning system called Argus (Xu et al. 2015) is based on mobile terminals. Argus extracts geometric constraints from the image data of crowdsourcing and maps these geometric constraints to RSS fingerprint space to distinguish the location information of similar fingerprints and reduce the fuzziness of fingerprints.

Problems and Challenges Faced by Crowdsensing

Crowdsensing provides a new way to realize in-depth sensing but also brings new challenges to the research of theory, technology, and application. It can be summarized as the following aspects:

- The efficient transmission of crowdsensing data. Many crowdsensing applications need to continuously collect sensory data and transmit it to the data center. Reporting the sensing data

based on the connection between the mobile cellular network and the Internet will consume too much user equipment power and data traffic, resulting in greater pressure for the mobile cellular network. Therefore, it is necessary to design energy-efficient data transmission methods, such as based on short-range wireless communication.

- Resource optimization of crowdsensing networks. Overcoming the resource constraints of mobile nodes in terms of energy, bandwidth, computing, etc. is the key to the practical application of crowdsensing networks. First, because the number of users and the availability of sensors can change dynamically over time, it is difficult to accurately model and predict energy and bandwidth requirements to complete a specified task. Second, it is necessary to consider how to select a valid subset of users and reasonably schedule the sensing and communication resources under resource constraints.
- Incentive mechanisms of crowdsensing networks. A crowdsensing application relies on a large number of ordinary users to participate, and users will consume their own equipment power, computing, storage, communication, and other resources and bear the threat of privacy leakage when participating in sensing. Therefore, a reasonable incentive mechanism is needed to pay for user participation. The cost is compensated to attract enough users to ensure the quality of the data collection required.
- User privacy protection. By collecting the sensing data which is related to a user's position, the accurate position information of the user can be obtained (Liu et al. 2019b). Through long-term monitoring and analysis of location information, the user's home and work addresses, daily activity scope, and common traffic routes can be found. Mining the sensing data of motion state sensor can obtain the sensitive information of a user's daily life habit, health status, and so on; combining with the environment sensing data, it can also get the user's situation at any time. Although the large-scale collection

of sensitive sensing data of mobile intelligent terminals can analyze and mine many valuable information, such as urban traffic congestion, road conditions, air quality, environmental pollution, infrastructure conditions, public services, etc., once these sensing data is leaked, it will seriously threaten the privacy of users, so it is necessary to take effective measures to protect users' privacy.

- Data and platform security. Since the mobile crowdsensing network gathers a large number of sensitive and private data and can mine a large number of information with great application value, it will greatly increase the risk of hacker attack and confidential data disclosure (Liu et al. 2019a). How to better protect the security of data and platform has become a prominent and urgent key issue.
- Data authenticity and integrity. To build an efficient and valuable mobile crowdsensing network depends on users to provide authentic, complete, and reliable sensing data. However, for collecting sensing data, malicious users may submit false sensing data; or, for sensing data transmission, they may lose some data, which may lead to mining out wrong information and making wrong decisions. Therefore, it is very important to take corresponding technical measures to ensure the authenticity and integrity of data.
- The balance between the improvement of sensing quality and efficiency and the optimal utilization of resources. Due to the large number of sensors and the dynamic change of performance, the task allocation and resource consumption prediction under resource constraints become more complex and difficult (Liu et al. 2017). We can collect the sensing data of different types of sensing devices to achieve the same sensing goal, but the sensing quality and resource consumption are different. For example, location information can be mined by collecting sensing data from GPS, Wi-Fi, and mobile cellular networks. GPS has the highest location accuracy and resource consumption. Therefore, the balance between sensing quality and resource consumption

needs to be solved. In addition, it is possible to execute multiple priorities and different resource consumption sensing tasks on the same sensing device at the same time.

Reference

- Ganti RK, Ye F, Lei H (2011) Mobile crowdsensing: current state and future challenges. *IEEE Commun Mag* 49(11):32–39
- Guo B, Chen H, Yu Z, Xie X, Huangfu S, Zhang D (2014) FlierMeet: a mobile crowdsensing system for cross-space public information reposting, tagging, and sharing. *IEEE Trans Mob Comput* 14(10):2020–2033
- Jiang Y, Xu X, Terlecky P, Abdelzaher T, Bar-Noy A, Govindan R (2013) Mediascope: selective on-demand media retrieval from mobile devices. In: *Proceedings of the IPSN*
- Kim S, Robson C, Zimmerman T, Pierce J, Haber EM (2011) Creek watch: pairing usefulness and usability for successful citizen science. In: *Proceedings of the CHI*
- Liu Y, Wu H, Xia Y, Wang Y, Li F, Yang P (2017) Optimal online data dissemination for resource constrained mobile opportunistic networks. *IEEE Trans Veh Technol* 66(6):5301–5315
- Liu Y, Quan W, Wang T, Wang Y (2018) Delay-constrained utility maximization for video Ads push in mobile opportunistic D2D networks. *IEEE Internet Things J* 5(5):4088–4099
- Liu Y, Hao L, Liu Z, Sharif K, Wang Y, Das SK (2019a) Mitigating interference via power control for two-tier femtocell networks: a hierarchical game approach. *IEEE Trans Veh Technol* 68(7):7194–7198
- Liu Y, Wang H, Peng M, Guan J, Xu J, Wang Y (2019b) DeePGA: a privacy-preserving data aggregation game in crowdsensing via deep reinforcement learning. *IEEE Internet Things J*. <https://doi.org/10.1109/JIOT.2019.2957400>
- Losey J, Allee L, Smyth R (2012) The lost ladybug project: citizen spotting surpasses scientist's surveys. *Am Entomol* 58(1):22–24
- Mathur S, Kaul S, Gruteser M, Trappe W (2009) ParkNet: a mobile sensor network for harvesting real time vehicular parking information. In: *Proceedings of the MobiHoc Workshop*
- Morishita S, Maenaka S, Nagata D, Tamai M, Yasumoto K, Fukukura T, Sato K (2015) SakuraSensor: quasi-realtime cherry-lined roads detection through participatory video sensing by cars. In: *Proceedings of the UbiComp*
- Popescu A, Grefenstette G, Moëllic PA (2008) Gazetiki: automatic creation of a geographical gazetteer. In: *Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries*
- Sehgal S, Kanhere SS, Chou CT (2008) Mobishop: using mobile phones for sharing consumer pricing informa-

- tion. In: Proceedings of the international conference on distributed computing in sensor systems
- White J, Thompson C, Turner H, Dougherty B, Schmidt DC (2011) Wreckwatch: automatic traffic accident detection and notification with smartphones. *Mob Netw Appl* 16(3):285
- Xu H, Yang Z, Zhou Z, Shangguan L, Yi K, Liu Y (2015) Enhancing Wifi-based localization with visual clues. In: Proceedings of the UbiComp
- Zhou P, Zheng Y, Li M (2012) How long to wait? Predicting bus arrival time with mobile phone based participatory sensing. In: Proceedings of the MobiSys

Mobile Data Offloading Techniques in WiFi-Cellular Networks

► WiFi Offloading in WiFi-Cellular Networks

Mobile Data Offloading via D2D-Assisted Communications in Wireless Networks

Yuan Wu^{1,2}, Kejie Ni¹, Xiaowei Yang¹, and Liping Qian¹

¹College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

²State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China

Technical Background

The past decades have witnessed the proliferation of smart wireless devices and mobile Internet services, yielding a rapid growth of the data-hungry mobile applications such as video surveillance, online gaming, and augmented/virtual reality. These applications have resulted in a tremendous traffic pressure in radio access networks (RANs) of wireless networks due to the limited radio resources. Conventional wireless infrastructures in RANs require to deliver the mobile users' data directly through the infrastructures, e.g.,

marco/micro base stations (BSs) and Wi-Fi access points (APs), which thus lead to a tremendous traffic pressure in RANs. For instance, due to the limited radio resources, mobile users might suffer from traffic congestion when delivering traffic through a same BS/AP at a hot spot and during the peak hours. Moreover, the interference/congestion may also lead to excessive energy consumptions of mobile users. Device-to-device (D2D) communication, an emerging paradigm in the fifth generation (5G) cellular networks, provides a promising efficient approach to accommodate the traffic demands. Instead of requiring the traffic to traverse through the cellular BS, D2D communications allow two nearby mobile stations (MSs) to perform a direct communication by using either the licensed inband spectrums or the out-band spectrums. Exploiting the close proximity between the MSs, D2D communications can bring lots of benefits, e.g., enhancing throughput, reducing traffic delay, and saving energy consumption. In particular, exploiting D2D communications to actively offload users' traffic has been envisioned as an efficient approach to relieve the traffic pressure in RANs and improve the efficiency in radio resource utilization (Li et al. 2014; Han et al. 2010; Wu et al. 2017, 2018a; Chuang and Lin 2012; Al-Kanj et al. 2014; Gao et al. 2014).

A concrete example of the D2D-assisted data offloading for content distribution in wireless network is shown in Fig. 1. Specifically, there exist a group of MSs, denoted by $\mathcal{U} = \{1, 2, \dots, U\}$, coexisting in a given area, and there exist a group of content blocks (CBs), denoted by $\mathcal{K} = \{1, 2, \dots, K\}$, available at the BS. We use L^k to denote the size of CB k . Each MS i is interested in receiving a subset of $\mathcal{K}$, and we use a U -by- K matrix $\mathbf{A}$ to denote the relationship between the MSs and the CBs, i.e., $A_{i,k} = 1$ if MS i is interested in obtaining CB k , and $A_{i,k} = 0$ otherwise. Set $\mathcal{C}_i$ denotes the set of the interested CBs which MS i wants to receive, i.e., $\mathcal{C}_i = \{k \in \mathcal{K} | A_{i,k} = 1, i \in \mathcal{U}\}$, and set Ω^k denotes the set of MSs which are interested in obtaining CB k , i.e., $\Omega^k = \{i \in \mathcal{U} | A_{i,k} = 1, k \in \mathcal{K}\}$. The group of MSs are in close proximity. To deliver CB k to all MSs in Ω^k , the D2D-assisted offloading mecha-

nism works as follows: The BS determines the transmission rate r^k of CB k , which is equivalent to its transmission duration $x^k = \frac{L^k}{r^k}$. The BS first sends part of data of CB k to some selected MS i , and meanwhile, MS i relays its received data via broadcasting to the other MSs in Ω^k . Let z_i^k denote the relaying duration of MS $i \in \Omega^k$ for CB k . There exists $\sum_{i \in \Omega^k} z_i^k \leq x^k$, the network scenario with $\mathcal{U} = \{1, 2, \dots, 8\}$, $\mathcal{K} = \{1, 2, 3\}$, and the detailed matrix **A**. Figure 1b, c, and d illustrate the process of distributing CB 1 to MSs 1, 4, 6, and 7 in a step-by-step manner, i.e., including Phase-I, Phase-II, and Phase-III, respectively. Figure 1b and c illustrate Phase-I and Phase-II of relaying through MS 1 and MS 4, respectively. In Phase-I, MS 1 sends its received data to MSs 4, 6, and 7. In Phase-II, MS 4 sends its received data to MSs 1, 6, and 7. Finally, taking into account the MSs' limited transmit powers and energy capacities, in Phase-III, the BS directly broadcasts the rest data of CB 1 to the MSs in Ω^1 for finishing the delivery of CB 1. Let z_B^1 denote the duration for the BS to perform this broadcasting. Then, there exists $z_B^1 + \sum_{i \in \Omega^1} z_i^1 \leq x^1$. Figure 1d shows the case of BS's broadcasting to all MSs in Phase-III.

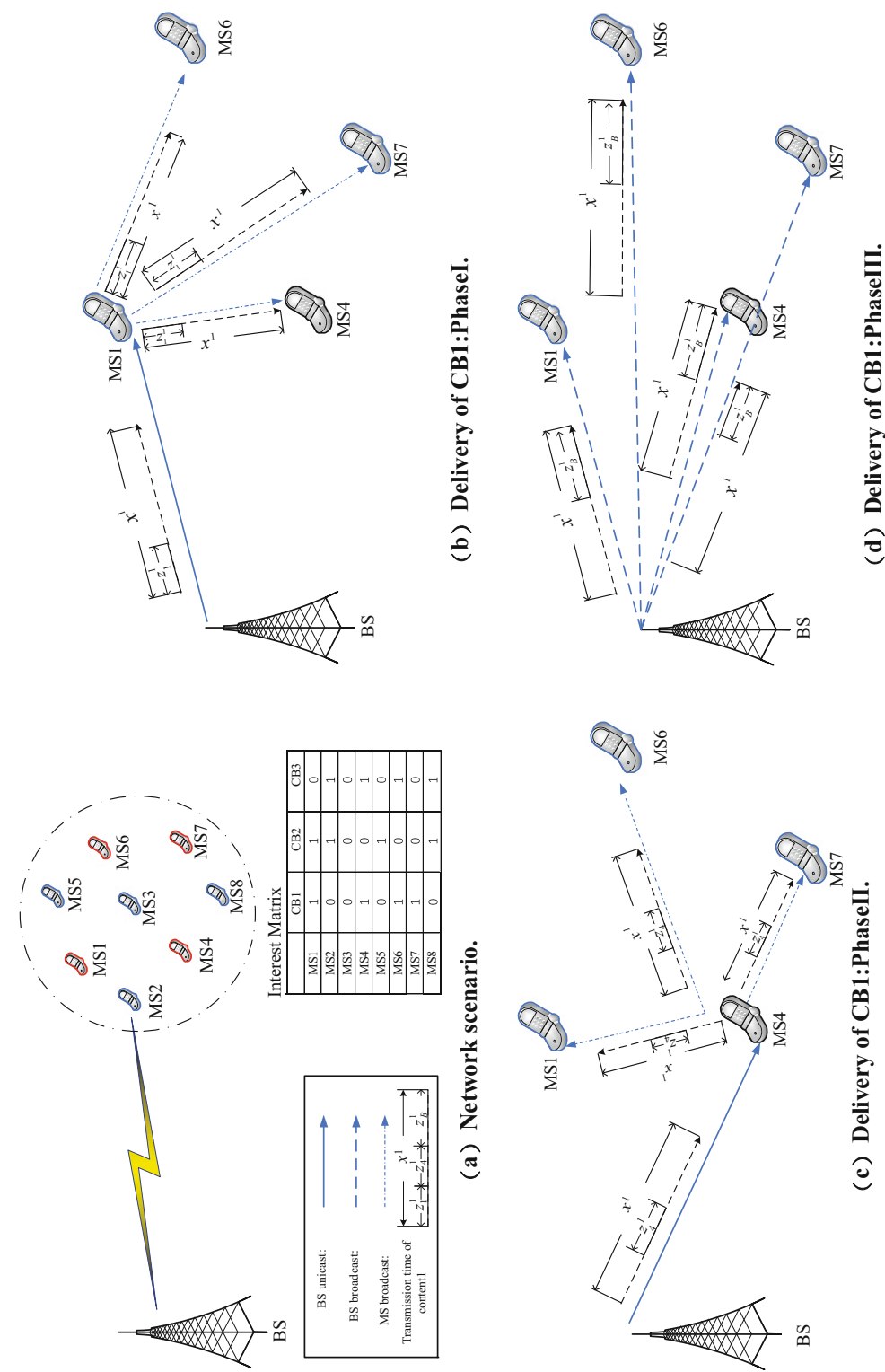
The D2D-assisted data offloading has attracted lots of interests, which can be in general categorized into two groups. The first group of studies focused on exploiting the D2D-assisted offloading for delay-tolerance applications. For instance, in (Li et al. 2014), a delay-tolerant network (DTN)-based offloading problem has been studied, which jointly takes into account the heterogeneous data traffics, users' different interests, and the limited storage resources. In (Han et al. 2010), accounting for the feature of mobile social networks, a target set selection problem for D2D-assisted information distribution has been investigated. In (Chuang and Lin 2012), exploiting the features of the encounter frequency and social communities, a community-based opportunistic D2D information dissemination scheme has been proposed, with the objectives of minimizing cellular traffic load and delivery time. The second group of studies focused on

exploiting the D2D-assisted offloading for delay-sensitive (real-time) applications. For instance, in (Al-Kanj et al. 2014), the authors investigated how to separate the mobile users into different groups and to select a leader of each group for distributing the contents.

Key Applications

The above mobile data offloading via D2D-assisted communications can be used for lots of applications. The following are two typical examples.

- *Vehicular networks*: The emerging vehicular communications/networks require data and message transmissions between the moving vehicles and infrastructures (e.g., roadside units (RSUs)) as well as the transmissions between the moving vehicles. Taking into account the mobility of vehicles and the delay requirement of data transmissions (e.g., for the safety-oriented applications), exploiting the D2D-assisted data offloading among the vehicular users provides an effective approach to improve the throughput and reliability of vehicular data transmissions.
- *Mobile edge computing*: Mobile edge computing (MEC), which enables mobile terminals to offload the computation tasks to nearby edge servers/terminals with sufficient computation resources, has been considered as a promising approach to implement the emerging computation-intensive mobile internet applications (e.g., augmented/virtual reality). However, MEC involves intensive data transmissions between the mobile terminals and the edge servers (e.g., a mobile user offloads its computation task to an edge server), which thus necessitate resource-efficient data transmissions with low latency and high throughput. The D2D-assisted data offloading has been envisioned as a promising approach to achieve this objective, and the recent advanced non-orthogonal multiple



Mobile Data Offloading via D2D-Assisted Communications in Wireless Networks, Fig. 1 An example of the D2D-assisted data offloading for content distribution

access (NOMA) (Wu et al. 2018b) also facilitates the massive D2D communications for data offloading.

References

- Al-Kanj L, Poor V, Dawy Z (2014) Optimal cellular offloading via device-to-device communication networks with fairness constraints. *IEEE Trans Wirel Commun* 13(8):4628–4643
- Chuang YJ, Lin KCJ (2012) Cellular traffic offloading through community based opportunistic dissemination. In: *Proceedings of 2012 IEEE wireless communications and networking conference*. pp 3188–3193
- Gao L, Iosifidis G, Huang J, Tassiulas L (2014) Hybrid data pricing for network-assisted user-provided connectivity. In: *Proceedings of 2014 IEEE INFOCOM*. pp 682–690
- Han B, Hui P, Kumar VS, Marathe M, Pei G, Srinivasan A (2010) Cellular traffic offloading through opportunistic communications: a case study. In: *Proceedings of the 5th ACM workshop on challenged networks*. pp 31–38
- Li Y, Qian M, Jin D, Hui P (2014) Multiple mobile data offloading through disruption tolerant networks. *IEEE Trans Mob Comput* 13(7):1579–1596
- Wu Y, Chen J, Qian L, Huang J, Shen X (2017) Energy-aware cooperative traffic offloading via device-to-device cooperations: an analytical approach. *IEEE Trans Mob Comput* 16(1):97–114
- Wu Y, Qian L, Mao H, Yang X, Shen X (2018b) Optimal power allocation and scheduling for non-orthogonal multiple access relay-assisted networks. *IEEE Trans Mob Comput*. <https://doi.org/10.1109/TMC.2018.2812722>
- Wu Y, Qian L, Zheng J, Zhou H, Shen X (2018a) Green-oriented traffic offloading through dual-connectivity in future heterogeneous small-cell networks. *IEEE Commun Mag* 56(5):140–147

Mobile Device Security and Privacy

► Mobile Security and Privacy

Mobile Edge Caching

► Mobile Edge Caching in HetNets

Mobile Edge Caching in HetNets

Xiuhua Li^{1,3}, Xiaofei Wang², and

Victor C. M. Leung³

¹Chongqing University, Chongqing, P. R. China

²School of Computer Science and Technology, Tianjin University, Jinnan, Tianjin, China

³The University of British Columbia, Vancouver, BC, Canada

Synonyms

HetNet; Mobile edge caching

Definition

Mobile edge caching refers to distributing content files (e.g., videos, audios, photos, application programs, and so on) from service providers over the Internet to caches that are deployed at the edges (e.g., mobile devices and base stations) of mobile networks, aiming at bringing content closer to mobile users in the distance of the network topology to deal with the challenge of the explosive growth in content requests from users in mobile networks. HetNet is short for heterogeneous network and is a form of radio access networks (RANs) with complex interoperation between macro cells and small cells, which consists of different types of base stations (BSs) such as femto BSs, pico BSs, micro BSs, and macro BSs.

Mobile edge caching in HetNets is the combination between the technique of mobile edge caching and the network architecture of HetNets, aiming to achieve their joint benefits in enhancing network performances. Due to the high disparity of content popularity (i.e., a small number of content files actually may attract a large amount of downloads), by deploying hierarchical collaborative caching at the edges of HetNets, the content delivery cascades can be optimized during the intermediate transmissions, while reduced content delivery delays can be also provided.

Historical Background

With the rapid advancement of mobile networks from 2G and 3G to current 4G, people's daily life has changed significantly, and people are increasingly enjoying online social activities on mobile devices (e.g., smartphones and electronic tablets). As a result, requests for various content files (e.g., videos, audios, photos, application programs, and so on) from mobile users are increasing at an explosive speed, which has become a serious issue of mobile network operators (MNOs). However, current RANs and backhaul networks cannot support these content requests effectively due to the scarcity of network resources and the limit of their network architectures (Wang et al. 2015). Thus, to deal with those issues, it is necessary to employ revolutionary schemes in network architectures and data transmissions toward the fifth-generation (i.e., 5G) mobile networks.

Mobile edge caching is regarded as an effective technique to reduce the reduplicated network traffic load in mobile networks. In particular, studies in (Cha et al. 2007; Chen et al. 2002) have shown that a large portion of the network traffic is caused by massive duplicated downloads of the same popular content. For instance, top 10% of videos in YouTube account for about 80% of all the views (Chen et al. 2002). Thus, with this technique, by caching popular content at the edges (e.g., mobile devices and BSs) of mobile networks, the requested content can be closer to mobile users in the distance of the network topology, and mobile users can directly access the content cached at the edges instead of downloading the content from SPs over the Internet via backhaul networks. Reduplicated transmissions from servers to clients are avoided, and most of the requests from users can be satisfied immediately in mobile networks. Consequently, mobile edge caching can effectively enhance the network performances, especially on offloading network traffic (Chen and Yang 2016; Li et al. 2016, 2017), reducing system costs (Zhi et al. 2016; Gregori et al. 2016), and improving the quality of service (QoS) or quality of experience (QoE) of mobile users (Golrezaei et al. 2012; Zhao et al.

2016; Ao and Psounis 2015; Hong and Choi 2016; Li et al. 2015).

Another effective approach is to introduce the architecture of HetNets, which consists of different types of BSs (such as femto BSs, pico BSs, micro BSs, and macro BSs) with different wireless coverages (Li et al. 2016, 2017). HetNets can greatly enhance wireless link quality between mobile users and BSs and thus improve network capacity.

Considering the great potentials of the above two techniques, it is beneficial to combine them together, i.e., mobile edge caching in HetNets, which can effectively address the above discussed issues of MNOs. Studies in (Wang et al. 2015; Chen and Yang 2016; Zhi et al. 2016; Gregori et al. 2016; Golrezaei et al. 2012; Zhao et al. 2016; Ao and Psounis 2015; Hong and Choi 2016; Li et al. 2015) focused on single-tier caching in either mobile devices or BSs in mobile networks. Studies in (Li et al. 2016, 2017; Yang et al. 2016; Jiang et al. 2017; Xu and Tao 2017) focused on hierarchical BS caching in HetNets, where the edge caching in mobile devices is not considered. Studies in (Rao et al. 2016; Wang et al. 2017) explored the mobile edge caching in both mobile devices and BSs in mobile networks.

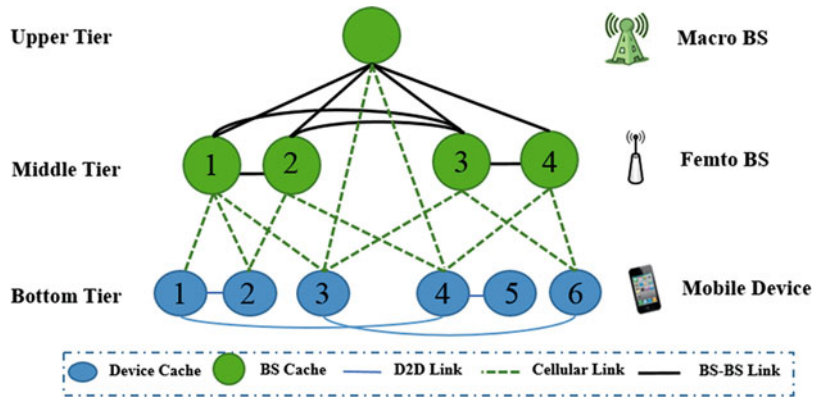
Foundations

To satisfy content requests from mobile users, mobile edge caching in HetNets needs to consider two phases, i.e., content placement phase and content delivery phase.

In the content placement phase, mobile edge caching in HetNets mainly deals with the following four issues:

- **Caching Topology** – To bring content closer to mobile users, it is important to decide where to cache in mobile networks. In HetNets, content can be cached at the edges that consist of mobile devices and BSs, which can form an either single-tier or hierarchical edge caching topology. Firstly, MNOs need to design the network architecture of HetNets,

Mobile Edge Caching in HetNets, Fig. 1 An illustration of caching topology



especially about how many tiers or types of BSs there are in HetNets as well as how different BSs connect. Secondly, MNOs need to consider whether mobile devices are able to cache content or not. Thirdly, MNOs need to consider which tiers of BSs are able to cache content or not. As a result, the mobile edge caching topology can be achieved, denoting the deployment of caches at the edges of HetNets.

As illustrated in Fig. 1, the caching topology of mobile edge caching in a HetNet is hierarchical and consists of three tiers of caching, i.e., bottom tier with six mobile device caches, middle tier with four femto BS caches, and upper tier with one macro BS cache. Here, the HetNet consists of two tiers of BSs, i.e., four femto BSs and one macro BS. All the BSs and mobile devices are able to cache content. Besides, the connection among BSs, among mobile devices, and between BSs and mobile devices is via BS-BS links, D2D links, and cellular links, respectively. In particular, mobile devices are possible to be connected with any tier of BSs. For instance, mobile device 3 is connected with femto BS 1 and femto BS 3 at the middle tier as well as the macro BS at the upper tier.

- **Content cacheability and storage format** – Mobile edge caching in HetNets aims to achieve a trade-off between network performances (e.g., network traffic load, system costs, and QoS/QoE) that are usually expensive to be improved and storage costs that are becoming much cheaper. However, the practical scale of content owned by SPs is growing rapidly, and thus it is impossible to cache all the content. Thus, it is important to decide what content to cache (i.e., content cacheability) taking content popularity into account. As practically captured in (Cha et al. 2007; Chen et al. 2002), only a small amount of popular content accounts for a large portion of content requests from mobile users, while a long tail of content remains unpopular. Besides, different types of content have different cacheability (Wang et al. 2015). For instance, among the content types, videos and photos have the highest revisit rate. Moreover, a large content file can be chunked into a series of original segments, and original segments can be encoded into an arbitrary number of packets to explore the diversity of cached content, while a given number of different coded packets can be collected together and decoded into the original segments for recovering the content file (Golrezaei et al. 2012; Yang et al. 2016). Thus, a content file can be cached by storing in three formats: (1) entire original content in a neither unchunked nor uncoded case, (2) original segments in a chunked but uncoded case, and (3) coded packets from the original segments in a chunked and coded case.
- **Caching policy** – Caching policies, deciding what to cache, how to cache, and when to release caches for what network objectives, are crucial to achieve the performance gains

of mobile edge caching. In particular, the optimization objectives of mobile edge caching in HetNets can be various for enhancing network performances, such as offloading network traffic, reducing system costs, improving mobile users' QoS/QoE, and so on. It is also important to estimate the gain of caching a content file by evaluating its current popularity, potential popularity, storage size, and locations of existing replicas in the network topology based on the system learning and analysis from mobile users' social behaviors and preferences (Li et al. 2016, 2017). Mobile edge caching policies can be operated in offline/online manners based on the practical network requirements. Offline caching (i.e., proactive) is relatively static with some prior network knowledge such as content requests and content popularity, while online caching (i.e., reactive caching) is relatively flexible according to the dynamics of network knowledge. Rather than employing traditional caching policies (e.g., least recently used (LRU), least frequently used (LFU), and first-in first-out (FIFO)), it is important and challenging to design proper cooperative mobile edge caching policies in HetNets to improve the network performances.

- **Operation time of caching** – According to the change of content popularity and operated manners of mobile edge caching, different types of content can be cached in different time periods. Online caching needs to cache content immediately or in a short time, while offline caching does not. For instance, in offline caching, short-lifetime popular news with short videos are updated every a few hours, while long-lifetime new movies and new music videos are, respectively, posted weekly and monthly (Li et al. 2017). In order to reduce the traffic load and avoid possible network traffic congestion especially in busy hours, content can be cached in off-peak hours (e.g., late night).

In the content delivery phase, each mobile user requests content based on its own preference. In

order to satisfy a user's content request, mobile edge caching in HetNets mainly deals with the following two issues:

- **Content request routing** – Content request routing shows the possible routes for delivering the requested content in the derived caching topology to a user before operating practical wireless transmissions of content in HetNets. Specifically, in the network, the whole process of content request routing can be summarized as:
 - If a user's requested content is locally cached in its mobile device, then the request can be satisfied locally.
 - Otherwise, the user can first find the content in caches of other users in close proximity and then establishes a device-to-device (D2D) link with a user where the content is available and finally fetches the content in a D2D manner (e.g., Wi-Fi Direct, or Bluetooth) (Gregori et al. 2016; Rao et al. 2016).
 - If not yet satisfied, the user has to be served by the associated BSs via cellular links. If the requested content is locally cached, then the associated BSs satisfy the request directly; otherwise, the associated BSs need to explore the cooperation possibility for fetching the content from other BSs where the content is available.
 - If not yet satisfied, then downloading the content directly from SPs over the Internet via backhaul networks is the last resort.
- **Wireless transmission of content** – After the content request routing is known, the requested content will be delivered to mobile users with wireless transmissions via either D2D links or cellular links. In terms of D2D links, they are established only when a pair of mobile users are in close proximity and willing to share the cached content in a D2D manner. Here, social behaviors and content preference of mobile users need to be analyzed and estimated in advance by system learning. In terms of cellular links, the associated BSs with noncooperation or

cooperation can transmit the requested content to the user by unicasting/multicasting based on the practical data transmission schemes in HetNets. In particular, resource allocation and scheduling for wireless transmissions of content via cellular links is necessary to enhance the network performances while satisfying the content requests.

Key Applications

The technique of mobile edge caching in HetNets can be utilized for general MNOs to deal with the explosive growth in content requests from mobile users, and thus there are also vendors that are manufacturing cache-enabled base station products, e.g., cache-enabled femtocell products and cache-enabled Wi-Fi routers. For companies of content delivery networks (CDNs), a new key trend is also the extension of their services into mobile edge networks, particularly for BSs with opened interfaces for MNOs and third-party content providers. Mobile edge caching in HetNets can also be utilized in 4G networks, but will be widely deployed in 5G networks.

Future Directions

From the evolution of mobile edge caching in HetNets, it appears that, in the future, emphasis will be given on the following research directions as follows:

- Big data-based large-scale online/offline optimization of content caching for MNOs
- Machine learning-based analysis and prediction on mobile users' social behaviors and preference for caching and prefetching optimization
- More rapid and secure collaborations among mobile devices and BSs in HetNets
- Pricing methodology for mobile edge caching in HetNets

Cross-References

- ▶ [Content-Centric Wireless Sensor Networks](#)
- ▶ [Resource Allocation](#)
- ▶ [Mobile Edge Computing](#)

References

- Ao WC, Psounis K (2015) Distributed caching and small cell cooperation for fast content delivery. In: Proceedings of the ACM MobiHoc, June 2015, pp 127–136
- Cha M, Kwak H, Rodriguez P, Ahn YY, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the ACM IMC, Oct 2007, pp 1–14
- Chen B, Yang C (2016) Caching policy optimization for D2D communications by learning user preference. In: Proceedings of the IEEE WCNC, May 2016, pp 1–6
- Chen Y, Qiu L, Chen W, Nguyen L, Katz R (2002) Clustering web content for efficient replication. In: Proceedings of the IEEE ICNP, Nov 2002, pp 165–174
- Golrezaei N, Shanmugam K, Dimakis AG, Molisch AF, Caire G (2012) FemtoCaching: wireless video content delivery through distributed caching helpers. In: Proceedings of the IEEE INFOCOM, Mar 2012, pp 1107–1115
- Gregori M, Vilardebó JG, Matamoros J, Gündüz D (2016) Wireless content caching for small cell and D2D networks. *IEEE J Sel Areas Commun* 34(5):1222–1234
- Hong JP, Choi W (2016) User prefix caching for average playback delay reduction in wireless video streaming. *IEEE Trans Wirel Commun* 15(1):377–388
- Jiang W, Feng G, Qin S (2017) Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. *IEEE Trans Mobile Comput* 16(5):1382–1393
- Li X, Wang X, Xiao S, Leung VCM (2015) Delay performance analysis of cooperative cell caching in future mobile networks. In: Proceedings of the IEEE ICC, June 2015, pp 5652–5657
- Li X, Wang X, Leung VCM (2016) Weighted network traffic offloading in cache-enabled heterogeneous networks. In: Proceedings of the IEEE ICC, May 2016, pp 1–6
- Li X, Wang X, Li K, Han Z, Leung VCM (2017) Collaborative multi-tier caching in heterogeneous networks: modeling, analysis, and design. *IEEE Trans Wirel Commun* 16(10):6926–6939
- Rao J, Feng H, Yang C, Chen Z, Xia B (2016) Optimal caching placement for D2D assisted wireless caching networks. In: Proceedings of the IEEE ICC, May 2016, pp 1–6
- Wang X, Li X, Leung VCM, Nasiopoulos P (2015) A framework of cooperative cell caching for the future mobile networks. In: Proceedings of the HICSS, Jan 2015, pp 5404–5413

- Wang W, Lan R, Gu J, Huang A, Shan H, Zhang Z (2017) Edge caching at base stations with device-to-device offloading. *IEEE Access* 5:6399–6410
- Xu X, Tao M (2017) Modeling, analysis, and optimization of coded caching in small-cell networks. *IEEE Trans Commun* 65(8):3415–3428
- Yang C, Yao Y, Chen Z, Xia B (2016) Analysis on cache-enabled wireless heterogeneous networks. *IEEE Trans Wirel Commun* 15(1):131–145
- Zhao Z, Peng M, Ding Z, Wang W, Poor HV (2016) Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks. *IEEE J Sel Areas Commun* 34(5):1207–1221
- Zhi W, Zhu K, Zhang Y, Zhang L (2016) Hierarchically social-aware incentivized caching for D2D communications. In: *Proceedings of the IEEE ICPDS*, Dec 2016, pp 316–323

Mobile Edge Computing

- [Computation Offloading in Mobile Edge Computing](#)
- [Data-Driven Edge Computing](#)
- [Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks](#)

Mobile Edge Computing: Low Latency and High Reliability

Zhiyuan Ren¹, Chen Chen¹, and Jun Fu²

¹Xidian University, Xi'an, Shaanxi, China

²Future Forum, Beijing, China

Synonyms

[Edge computing](#); [Mobile cloud computing](#)

Definition

Mobile edge computing (MEC) has been proposed in recent years for offloading computation tasks from user equipment (UE) to the network edge to break hardware limitations and resource constraints at UE. Through migrating the computation, storage, and servicing capability to the

edge of network, MEC can deploy application, service, and content locally in a distributed way. To a certain extent, it will satisfy the critical low latency and high reliability requirements for future eMBB, uRLLC, mMTC scenarios.

Historical Background

The European Telecommunications Standards Institute (ETSI) proposed mobile edge computing (MEC) in 2014, which is a technology based on 5G evolution architecture and deeply integrates the base station (BS) and Internet services. This technology is highly concerned by operators and equipment vendors, i.e., Huawei conducted joint tests with major operators and believes that MEC is one of the key technologies for coordinating user needs and network capabilities, which will help customers to achieve the goal of intelligent manufacturing, such as production, maintenance, management, inspection, and security. In 2016, the Edge Computing Consortium (ECC) was formally established, which would further promote the development of MEC technology.

At present, researches on MEC mainly focus on the computing offloading, energy efficiency, edge storage, low-latency guarantee, etc.

Computing offloading means that the terminal device with limited computing capability transmits the heavy computing task to the nearby MEC servers, and the MEC server replaces the terminal to complete the computing task. After the computing task is over, the MEC server returns the result to the terminal device. Chen et al. (2016) proposed a game theoretic approach for the computation offloading decision making problem among multiple-users in multichannel environments. To solve the problem of computing offloading in the Internet of Vehicles, Zhang et al. (2016) proposed a cloud-based mobile edge computing offloading architecture and designed an efficient allocation schedule of computing resources. Under the latency constraints of computing tasks, the resource utilization can be improved. Swaroop Nunna et al. (2015) proposed the concept of constructing next-generation collaborative architecture based on

MEC server under 5G network. It is located at the edge of mobile network and cooperates with the underlying communication network, which provides a potential architecture solution for the environment-aware collaboration of critical mission. Jararweh et al. (2016) proposed a software-defined system based on edge computing (SDMEC). Through the collaborative work of SD-Network, SD-Compute, SD-Storage, and SD-Security, the capabilities of cloud are expanded to the edge of the network, which satisfied the requirements of some applications, i.e., traffic regulation, content sharing, mobile gaming, etc.

In the view of the bright prospects of MEC, academics and industries have already carried out research on edge computing and have made some progress. However, the MEC technology is still immature, and specific implementation plans and related technical standards are still evolving.

Foundations

MEC offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the network, as Fig. 1 shows. Through migrating the services which requires high complexity and high-energy consumption to the MEC server, the smart mobile terminals could solve the problem of limited capability in computation, storage, and battery. In additional, MEC can effectively reduce the demand of network bandwidth and the increasing transmission pressure of the core network with local offloading, caching, and deployment. To satisfy the future requirements of eMBB (enhanced mobile broadband), uRLLC (ultra-reliable low-latency communication), and mMTC (massive machine type communication), the following key technologies under MEC have been studied:

Traffic offloading. Traffic offloading is one of the basic features in the MEC. By leading a part of traffic to local network (campus network or enterprise network), traffic offloading can greatly reduce the stress of core network bandwidth and decrease the transmission delay.

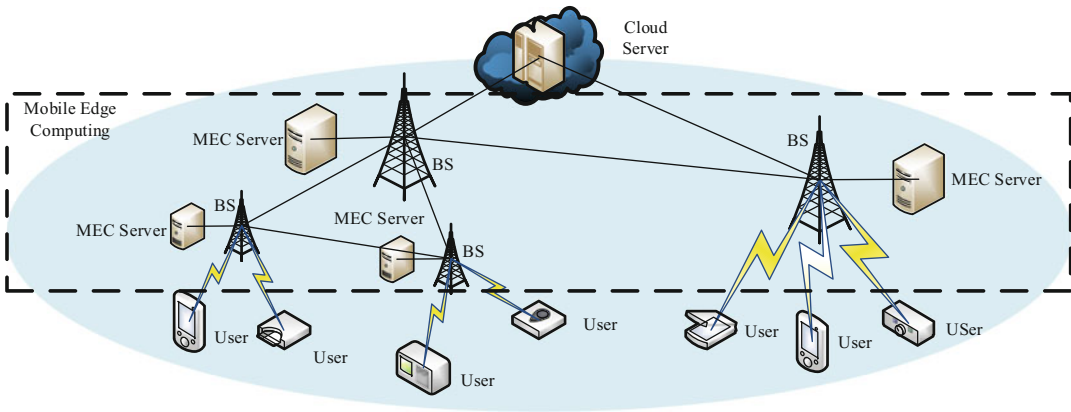
The future network will be a merging network, which includes 4G, 5G, Wi-Fi, etc. Therefore, the future deployment of applications need to consider how to realize the local traffic offloading technology based on MEC in 4G networks. In addition, the development of the local traffic offloading technology in MEC needs to further clarify the commercial model.

Cache and acceleration. MEC can deploy the content closer to the end user. Thus, data can be directly obtained from local server, which would greatly reduce the service access delay. Besides, several questions need to be considered when caching is used in MEC:

1. According to the MEC typical service scenarios, and the existing caching models (local DNS, redirection, transparent proxy), MEC needs to choose appropriate caching model.
2. MEC needs to design efficient algorithm to optimize the cache performance or make a trade-off between its deployment place and the hitting ratio.
3. Cache channel from MEC to remote server needs to be optimized to support cache acceleration scheme based on MEC.
4. To avoid the same video content repeatedly download due to different resolution (different bit rate), MEC needs to consider how to cache the same video content (HD) version and regenerate the content for different resolution (rate) to provide service for different kinds of terminals.

Mobility management. Under the MEC environment, to provide service continuity, new mobility management needs to be deployed at the edge of the network. The mobility management is expected to handle three scenarios including the change of data forwarding path caused by UE mobility, application data migration due to load balance, and interaction between MEC system and other systems when the UE moves in/out MEC service area.

MEC-based computing offloading. MEC-based computation offloading is a technology which leverages the edge devices near terminals to process task. By using a series of network



Mobile Edge Computing: Low Latency and High Reliability, Fig. 1 The architecture of MEC

equipment near the BS, the delay sensitive service should be offloaded to the edge devices for rapidly response and lower latency. Moreover, besides eMBB scenario, the 5G network contains lots of low-latency high-reliability services, such as V2X, intelligent manufacturing, etc. These applications require many sensors or terminals for information collection, then transmit the information to the network side for cooperative process. Although the computing capability of single edge device is finite, the network has large number of edge devices which means the distributed computing can be used in this scenario. Therefore, the computing capacity of all the edge devices could be combined, which could form a distributed edge computing pool to remarkably reduce the computing pressure of cloud platform, ensure the QoS of communication, and improve the stability and reliability of the network.

MEC security. The MEC server is a brand-new type entity in the mobile network. While connecting to several mobile network entities including Operation Administration and Maintenance (OAM) system or Lawful Interception, etc., it also connects to the third-party application server and even accommodates the third-party application. It is noted that current security techniques including IP sec, TLS, firewall, etc. could secure these connections to block the attack to the MEC server, mobile system, and the third-party server. However, these solutions are not

safe enough for the new security issues with the introduction of MEC. It is necessary to setup a unified, credible security evaluation system to evaluate the security of the applications running in the MEC system, which will only guarantee the trustable third-party application in the server. Meanwhile, the security aspects of the interface between the application and the network should also be considered for safely communication, e.g., the unauthorized invocation. There are also new business opportunities for operators to provide the security service. The mobile operator could provide the service to shield the application in the MEC server from the malicious attack and check the security bugs of the third-party apps.

Key Applications

According to the definition of the ETSI Industry Specification Group (ISG), MEC offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the network. This makes it possible to deploy the application, service, and content closer to the UE in a distributed manner. The MEC application focuses on the scenarios of low latency and high bandwidth. According to different service characteristics, the MEC applications can be divided into the following two categories: local service and vertical industry. The local service

applications mainly consist of campus network and AR/VR service, etc. And the vertical industry applications mainly include the V2X and industrial Internet scenarios. The following are some typical applications of the local service and vertical industry applications.

Local Service

Typical Application 1: Local Video Surveillance

In the scenario of local video surveillance, the traffic volume of video surveillance especially HD video is very huge. For example, the uplink bandwidth is up to 16Mbps for 4K HD video. If all data must be sent to the central cloud, it will consume large backhaul bandwidth. On the other hand, the video needs to be processed and analyzed timely to handle the emergency situation. The MEC can satisfy the user requirement by offloading the traffic and processing the data locally to provide a better user experience. Besides video surveillance, many applications like sports event live require low latency and high bandwidth, and MEC can significantly enhance the user experience and reduce the operator network load in such environment.

Typical Application 2: Augmented Reality

There exist many scenarios for AR, such as museum, city memorial, sports event, and concert, which can enhance the user experience remarkably. MEC can achieve efficient office automation assistance, local resource access, and internal communication in these applications and can provide the consumer better experience and the ability to access local service.

Vertical Industry

Typical Application 1: Autonomous Driving

The autonomous driving vehicles generate large amounts of data. For example, one autonomous driving vehicle can generate and process TB level data per hour, which requires high bandwidth for data delivery and needs immediate processing the data and delivers the result to the vehicles with ultra-low latency. The latency requirement for assistant driving is 20-100 ms,

but autonomous driving requires 3 ms. MEC is one of the technologies for realizing low-latency traffic in mobile networks. There are two types of MEC deployments for Internet of Vehicles (IoV). The first one needs building an edge cloud system at the base station side, and the other one uses strong computing capability BS to provide low-latency services for mobile terminals.

Typical Application 2: Industrial Internet

Industry Internet of Things (IIoT) enables intelligent manufacturing for future industry, whose production data would be generated locally. The data processing and the result feedback need ultra-low latency which may be lower than 1 ms. Thus, the processing and computing can only be finished locally due to the privacy requirement. MEC can satisfy this requirement, which can also complete the coordination of different production processes, such as Drone and Robot.

Further Directions

MEC is a new network architecture concept, which has many new features comparing to the existing 3G/4G cellular systems with better latency and reliability performance. Therefore, there are many research challenges and opportunities in the future research works. The security of the MEC, the integration of the computing, storage and communications in the MEC architecture, the online caching in the MEC, etc. should be further researched.

Cross-References

- [Cloud Computing](#)
- [Computation Offloading in Mobile Edge Computing](#)
- [Mobile Edge Computing](#)
- [Mobility Management](#)

References

- Chen X, Jiao L, Li WZ, Fu XM (2016) Efficient multi-user computation offloading for mobile-edge cloud computing. *IEEE/ACM Trans Netw* 24(5):2795–2808
- Jararweh Y, Doulat A, Darabseh A et al (2016) SDMEC: Software defined system for mobile edge computing. Paper presented at IEEE international conference on cloud engineering workshop, Berlin, 4–8 Apr 2016
- Nunna S, Kousaridas A, Ibrahim M (2015) Enabling real-time context-aware collaboration through 5G and mobile edge computing. Paper presented at 12th international conference on information technology – new generations, Las Vegas, 13–15 Apr 2015
- Zhang K, Mao YM, Leng SP, Vinel A et al (2016) Delay constrained offloading for Mobile Edge Computing in cloud-enabled vehicular networks. Paper presented at 8th international workshop on resilient networks design and modeling, Halmstad, 13–15 Sept 2016

Mobile IP

Charles Perkins
Huawei Technologies, Santa Clara, CA, USA

Synonyms

MIP; MIPv4; MIPv6

Definition

Mobile IP (MIP) is a mobility management protocol developed within the IETF (Perkins 2002; Perkins et al. 2011). It is often also used to refer to one of, or a suite of, related protocols such as Proxy MIP, Mobile IPv6, Hierarchical Mobile IP, Fast Mobile IP, and numerous extensions for these.

Historical Background

The development of Mobile IP started in the early 1990s following on work done at Columbia University, IBM T.J. Watson Research, Carnegie Mellon University, and a number of other research groups. Further development

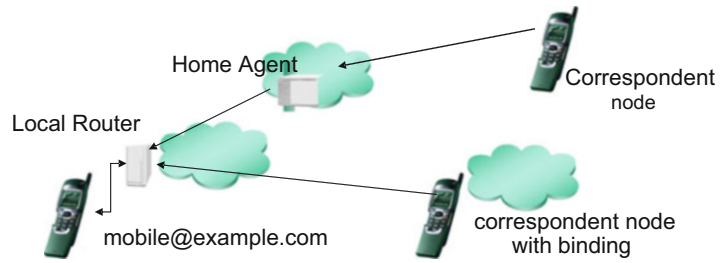
has continued in research departments around the world for many years, investigating the effects of Mobile IP on TCP, performance under high packet loss, hierarchical routing schemes, distributed mobility, interactions with BGP, etc.

The main problem solved by Mobile IP was to loosen IP's tight binding between the identity and the location of a mobile device. Since the beginning of the Internet, network-addressable devices had been known as stationary computers, and so routing data to the computer running an application really amounted to using routing prefix of the IP address to locate the network on which the computer was located. Put another way, once data arrived at the location determined by the IP address, there was not any question about the identity of the networked computer that would receive it. With the arrival of smaller mobile computers in the late 1980s, this assumption was no longer valid.

Protocol Operation

Mobile IP solves this problem by associating two IP addresses with a mobile device. One is the locator, and this IP address changes when the device moves to a new network. The other IP address is the device's identity, which does not change. The protocol itself is designed as an extension to IP, which binds (associates) the two IP addresses at a router on the device's "home network." The home network is the network to which data will be routed based on the IP address which identifies the mobile device. Once data arrive there, the router on the home network instead routes the data to the IP address which is the locator of the mobile device. The router that does this is called the "home agent." The locator IP address is routable to the network (called the visited network) currently visited by the mobile device. The association between the identifying IP address and the locating IP address is called a "binding," and as the locator IP address changes, the binding changes to reflect the updated address.

The process of updating the binding is the central design point for Mobile IP – and, indeed,

Mobile IP, Fig. 1

a similar requirement exists for practically every mobility management protocol whether or not developed in the IETF. Naturally, there are many details. Binding updates have to be done securely, to prevent unauthorized network devices from binding the MN's identifier IP address to a false locator IP address. There are many ways to prevent unauthorized tampering with the true locator IP address, and this is an area of protocol design that is still evolving in the Internet today. Another design area is related to the mechanism by which the home agent redirects traffic from the home network to the current location of the mobile device. This usually involves tunneling (encapsulating) the data once it arrives at the home network by putting on a new IP header that will cause the data to be routed to the locator IP address.

Many tunneling protocols exist in the IETF, and probably every one of them could be used for the purpose of redirecting traffic from the identifier IP address to the locator IP address, which in Mobile IP documents is typically called the "care-of address." Moreover, it is also possible to insert into a data packet a "source route" which supplies a list of intermediate routing points that can be followed to ensure the ultimate delivery of the packet. Tunneling and source routing are more or less equivalent, but tunneling has had more design work in the IETF and so has generally been favored for redirecting traffic as with Mobile IP. Lately, however, source routing is receiving renewed interest in the IETF (check the segment routing WG Fig. 1).

In the figure, a mobile device (mobile@example.com) has established a point of attachment with the Internet by way of a local router. For the rest of this entry, the mobile

device will be called a mobile node, abbreviated MN, as is usually done in IETF documents. Any other Internet device that communicates with MN will be called a Correspondent Node (the name often used in Mobile IP documents for the other communication endpoint for some traffic to/from MN). Traffic from the Internet can reach the mobile node (MN) by arriving at the Home Network, where the packets will be redirected (tunneled) to the local router currently serving the mobile node.

As a general observation, mobility management systems universally require some way for packets to be redirected to the current point of attachment of a mobile node. Determining the address for the redirection requires a lookup operation somewhere in the Internet. For business reasons, wireless operators have demanded that this lookup operation occurs within the operator's network, but this is not mandated by Mobile IP. The exact architectural placement of the lookup operation has a crucial effect on the performance of mobility management. Mobile IP is unique among mobility management systems in that the lookup operation occurs naturally on the routing path of the traffic to the mobile node, as determined by the mobile node's home address. All other systems have to insert some architectural augmentation for this to happen, which can easily be the source of bugs, performance loss, and other problems.

Packets routed to and from MN, which are redirected from the home network, experience additional delay because of the longer routing path. Depending upon local requirements in the network visited by MN, it is also likely that packets from the mobile node will have to be tunneled back to the home network. However, if a

correspondent node is able to receive information about the IP address of the mobile node in the visited network, then that correspondent node can use the information to route packets directly to and from the visited network. Any method by which the correspondent node is able to obtain the mobile node's care-of address is a kind of route optimization; various such methods have been proposed for MIPv4 and MIPv6.

Mobile IP for IPv4 (MIPv4) (Perkins 2002) and Mobile IP for IPv6 (MIPv6) (Perkins et al. 2011) are very similar in general operation, but there are important differences. The first distinction is that MIPv4 specifies the behavior of the local router, which for MIPv4 is called a Foreign Agent (FA). The Foreign Agent can issue periodic beacons to notify mobile devices that the FA is offering mobility services, and also provide a local IP address (i.e., a locator IP address). Once the mobile device MN identifies a Foreign Agent, then MN can use a (topologically correct) locator IP address as the local endpoint for traffic which will be tunneled from the Home Network. Thus, equipped MN supplies the locator IP address to the Home Agent and instructs the Home Agent to intercept traffic destined for the MN. If the tunnel is established between the Home Agent and the Foreign Agent, then the FA decapsulates packets and delivers them to the MN. Otherwise, the tunnel can be established between the Home Agent directly with the MN, and the FA has only to route the encapsulated packets as usual.

MIPv6, on the other hand, does not need to specify a Foreign Agent. By the design of IPv6, MN can get a locator IP address either by way of Router Advertisement or by way of DHCP (depending on local network administrative requirements); local IPv6 routers can be expected to support DHCP. Also, MIPv6 enables a special kind of routing header that causes packets to be delivered to MN's care-of address, avoiding the extra overhead of IPv6-within-IPv6 encapsulation.

As noted above, binding updates have to be performed securely. When Mobile IPv4 was designed, there was no IPsec. As a result, various Mobile IPv4-specific mechanisms were designed. These are still in use with MIPv4 and have

not been reported as broken or suffering from vulnerabilities that have plagued many other security protocols. The mip4 working group of the IETF even designed a key management and allocation protocol which enabled establishment of session keys for mobility management, based on an underlying security association between the home agent and the mobile node.

The security design of Mobile IPv6 was largely designed to utilize IPv6's security design (IPsec). IPsec offers authentication by insertion of a specially encrypted signature to guarantee that tampering can be detected. For further privacy, IPsec also offers payload encryption although this is not demanded for the safety of Mobile IPv6 signaling. Key management and refresh is also handled by standard IPsec methods.

Related Protocols

PMIP: At the time that Mobile IP was designed (late 1980s and early 1990s), IPv6 did not exist, and telecommunications was often considered synonymous with the international telephone system. Internet communication systems were mainly built on Unix systems running TCP/IP, and voice-over-IP (VOIP) did not exist for all practical purposes. In fact, VOIP was made illegal in some countries. As a consequence, the Internet was a small fraction of the total volume of communications. Internet applications typically did not have real-time requirements and ran on devices that did not move. As a further consequence, Mobile IP was not offered by the operating systems of the time that were dominant on laptop computers, which were the only ones that could move (albeit slowly). As the importance of Internet communications grew, network operators preferred to manage mobile devices without updating them to run Mobile IP. Instead, a network-proxy approach was desired which would restrict mobility management to the domain of a particular operator and still allow devices to be mobile within that domain even though the mobile devices did not have Mobile

IP available. This proxy approach is called Proxy Mobile IP (PMIP) (Gundavelli et al. 2008).

HMIP: An important variation of Mobile IP is “Hierarchical Mobile IP” (HMIP) (Soliman et al. 2008). HMIP was designed to separate the handling of local mobility (within one operator domain) from mobility between domains. The basic idea is to run local mobility management signaling to a local home agent as long as possible, and only transmit control messages to the (possibly distant) home network when the mobile device makes a new point of attachment outside of the domain which it had been visiting.

FMIP: A more sophisticated approach for reducing the latency penalty associated with mobility signaling across the larger Internet during handover has been designed and is known by the name of Fast Mobile IP (FMIP) (Koodli 2009). The basic idea is to enable neighboring IP access points to cooperate with each other by sharing information about the mobile node (MN). As MN moves from the network of one router (say, R1) to the network of another router (say, R2), R1 can send information to R2 about the mobile node, including its home address and its current locator IP address, temporary security credentials, multicast subscriptions, and other information (called “context”) about the mobile node to speed up the handover. Then, while packets in flight are still arriving at router R1, those packets can be further redirected to R2, the router serving MN at its new point of attachment. The signaling between the routers exchanges the context and sets up the temporary tunnel between the routers. The tunnel is configured to remain active for a long enough time so that the home agent can update its binding for MN to reflect the locator IP address associated with R2’s network.

Route optimization: As mentioned previously, route optimization techniques have been investigated for Mobile IP, so that a mobile node’s packets could be exchanged directly with a correspondent node without traveling to and from the home agent. One of the important contributions of Mobile IPv6 was to integrate a route optimization mechanism directly with the base protocol (Arkko et al. 2007). The route optimization protocol messages have the simple goal of

establishing a binding (as before, an association between the MN’s home address and its care-of address) at the correspondent node that can be used for tunneling. Then the correspondent node can use that binding to deliver packets directly to the mobile node’s care-of address. This was ground-breaking because it was the first time that a secure operation was designed to work across such a potentially large population of Internet devices (namely, the many possible correspondent nodes that might interact with a mobile node). The observation that enabled such a scalable approach to security was that the binding update can be authorized by assuring that the new care-of address belongs to the same mobile node with which a correspondent node had already been exchanging packets. This is a significantly weaker requirement than verifying the identity of a mobile node, and yet perfectly strong enough to prevent hijacking the mobile node’s traffic with its correspondent node.

Future Directions

Mobile IP and its related protocols have provided many innovations and a scalable approach to wide-area mobility management in the Internet. For various reasons, Mobile IP as specified within the IETF remains sparsely deployed in today’s wireless operator networks, but it has received a great deal of attention and has been influential in the evolution of modern telecommunication networks. With the advent of 5G networks, Mobile IP and its descendants within the [dmm] working group are seen as viable candidates for some 5G mobility management needs. With multigigabit wireless technologies already available, the high speed and architectural simplicity of Mobile IP become more attractive, as well as its inherent ability to serve all applications without requiring per-application mobility support.

Cross-References

- [Distributed IP Mobility Management](#)
- [Network-layer Mobility Management](#)
- [Proxy Mobile IPv6](#)

References

- Arkko J, Vogt C, Haddad W (2007) Enhanced route optimization for Mobile IPv6. RFC Editor
- Gundavelli S, Leung K, Devarapalli V et al (2008) Proxy Mobile IPv6. RFC Editor
- Koodli R (2009) Mobile IPv6 fast handovers. RFC Editor
- Perkins C (2002) IP mobility support for IPv4. RFC Editor
- Perkins C, Johnson D, Arkko J (2011) Mobility support in IPv6. RFC Editor
- Soliman H, Castelluccia C, ElMalki K, Bellier L (2008) Hierarchical Mobile IPv6 (HMIPv6) mobility management. RFC Editor

Mobile Networks

- [Capacity of Wireless Ad Hoc Networks](#)

Mobile Node

- [Proxy Mobile IPv6](#)

Mobile Security

- [Mobile Security and Privacy](#)

Mobile Security and Privacy

Feng Lin
Department of Computer Science and
Engineering, University of Colorado Denver,
Denver, CO, USA

Synonyms

[Mobile device security and privacy](#); [Mobile security](#); [Smartphone security and privacy](#)

Definitions

The protection of smartphones, tablets, other portable computing devices, and the networks

they connect to, from threats and vulnerabilities associated with wireless computing

Historical Background

Mobile security and privacy research has brought public attentions since the prevalence of cellphone and cellular networks. Notare et al. (1999) proposed a distributed security management system for telecommunications networks against cloned cellphones (same number and series of a genuine phone) in 1999. Thereafter, research in this area can be divided into two eras. The first era focuses on normal cellphones and voice services, such as cellphone authentication (Manabe et al. 2009), cellphone identification (Celiktutan et al. 2007), and cellphone cloning issue (Singh et al. 2007).

Since the debut of iPhone in 2007 and later the Android phones, research in mobile security and privacy is gradually shifting from cellphones to smartphones, which enriches more diverse application scenarios concentrating on mobile apps and mobile networks. The potential threats in mobiles may be caused by social engineering, compromised devices, malware, web browser or OS vulnerability, vulnerable application, and data interception. Especially, repackaging is one type of common attacks for smartphones targeting mobile apps, in which an attacker could modify a legitimate app to include malicious code and publish on third-party app stores. Correspondingly, the countermeasures have been developed to protect the mobile security and privacy, for example, multifactor authentication scheme that could avoid attacker easily guessing password or brute-force attack; detecting malware behaviors could play a significant role in reducing attack risks.

Different from traditional computer systems, the primary design objectives for mobile devices are low power consumption and portability. Hence, there are five key aspects of device- and application-level security features of mobile platform that distinguish mobile security from conventional computer security (La Polla et al. 2013): mobility, strong personalization, strong

connectivity, technology convergence, and reduced capabilities.

Mobile Vulnerability

- **Web-based services**

The mobile web is the most used platform other than mobile operating system (OS) platform. Similar to traditional web services on personal computers, mobile web applications that employ lightweight pages also suffer from security threats, such as phishing scams, drive-by downloads, browser exploits, cross-site scripting (XSS), SQL injection, etc. On the other hand, cyberattacks can exploit software vulnerabilities via mobile web browser to compromise mobile devices (Coursen 2007), such as PDF reader or image viewer.

- **Wireless networks**

Mobile devices can provide Internet connection and device-to-device communications at any time and in any place thanks to the wireless networks technologies, including Wi-Fi, cellular networks, Bluetooth, and near-field communication (NFC). Though these wireless networks provide convenience for users, using them in a public setting may cause potential risks. For example, Wi-Fi sniffing happens frequently with free Wi-Fi networks in airports, malls, and other public places when the mobile device is connecting to a malicious wireless access point (Beyah and Venkataraman 2011) and, hence, causes mobile devices' data exposure to attackers including account and personal information (Sujithra and Padmavathi 2012). Mobile networks may also suffer from overbilling attacks in which additional fees, such as fees of data traffic a user never used, are charged to the victim's accounts and transferred to the attacker's wallet.

- **Mobile malware**

As of the first quarter of 2018, more than 7 million of mobile apps are on the markets,

where Android users were able to choose between 3.8 million apps and Apple users can choose 2 million available apps in Apple App Store (Statista 2018). Among such huge mobile apps, there are many individual-developed third-party apps with malicious intent. Once those malicious apps, which may contain virus, Trojan, or botnet, are installed and gain access to the mobile devices, they could cause serious security breaches and incidents. As reported in Bradley (2011), more than 50 apps have been found to be infected by an Android malware called DroidDream in the official Google Play Store, which can stealthily gain root access to the device and further download additional malicious programs without user's knowledge and permission. As an example of smart malware, a multifarious malware for iOS devices, called iSAM (Damopoulos et al. 2011), has recently been designed, which integrates several typical malicious features of malware such as collecting confidential data stealthily and denial of application and network services, sending a large number of malicious SMS.

- **Weak authentication**

Most contemporary mobile devices use PIN, graphical pattern, or biometrics-based authentication such as fingerprint scan or face recognition. Those authentication methods are either vulnerable to malicious hacking, or the credentials are easily to be disclosed to others intentionally or voluntarily. PIN and pattern can suffer from *smudge attack* (Aviv et al. 2010), in which the password can be inferred by the smudge left on the screen surface when your fingertip touch on the screen. The brute-force attacks could be another threat to the PIN-based password. Though the biometrics-based authentication is superior because of the uniqueness of individuals, the biometric credentials are easy to be duplicated or counterfeited as anyone can obtain your fingerprint from a cup that you have hold.

What is even worse is that an attacker can obtain such biometric information from a distance. Chaos Computer Club announced that one of its members had been able to replicate the fingerprint of German Minister of Defense Ursula von der Leyen, using only photographs taken of her finger (CCC 2014).

Goal of Attacks

- **Denial of service**

Due to the portability and small-sized design, mobile devices usually have capability-limited hardware and battery, which restrict the computing, transmission, and power supply of the mobile device. Therefore, denial of service (DoS) attack is prone to target these shortages of mobile devices to disable one service, reduce the capability, or even make the whole device unusable by utilizing different attack techniques such as broadcasting highly malicious traffic stream, sending huge messages, or increasing power consumption. Battery power exhaustion attack is one of those DoS attacks that drains the battery faster than usual to forcefully shut down the device. The power of the device runs out up to 22 times faster than its normal condition by exploiting the wireless networks protocol vulnerabilities (Racic et al. 2006). The *water torture* attack (Johnston and Walker 2004) is another example of battery exhaustion attack carried out at the physical (PHY) layer that forces the device to send bogus frames. In addition, low-end mobile devices can be forced to shut down by receiving huge amount of SMS.

- **Privacy leakage**

Many mobile apps collect user data without users' permission such as address book or location information. A social media app was found to download users' full address books including names, phone numbers, and email addresses without users' knowledge or consent when enabling "find friends" feature. It is reported that mobile apps on

iPhone and iPad send personal information to advertising networks without the user consent too (Whitney 2010). Another most concerned privacy leakage is the location-related information. With the social networks apps becoming prevalent in our daily life, location-specific content have become more accessible and personalized to users' own context. Location tracking attack refers to attacker attempts to reveal the mobile device's location over time through examining their communications or hacking GPS information. Furthermore, if the business information is stored on the device, such invasions not only hurt the user's privacy but also increase the likelihood of security compromise to enterprise security.

- **Sniffing sensor information**

Comparing to traditional cellphones that are mainly used for communications, modern smartphone is a sensor-rich device that can provide extended functions equipped with multiple sensors, e.g., camera, microphone, GPS, inertial motion unit (IMU), compass, step recorder, etc. Once the attacker can access those sensors without authorization, all the user's actions will be sniffed and recorded. The stealthy video capture spyware can secretly start the built-in camera to record the private video, and it consumes little power that will not draw any attention from the user (Xu et al. 2009). Another example of compromised sensor is the microphone; *Soundcomber* (Schlegel et al. 2011) managed to extract private data, such as the touch sound of PIN or phone number, from the microphone.

M

Secure Protection Guidelines

- **Advanced authentication**

One-time validation of a user's identity, referred to as static authentication, has shown its vulnerability to attacks. Specifically, malicious adversaries may access the mobile

device that has been logged in by an authentic user when the authentic user is not nearby. Continuous authentication represents a new security mechanism which continuously monitors the user's trait and uses it as a basis to reauthenticate periodically throughout the login session. Hence, it has been adopted to overcome the limitations of traditional static authentication. These continuously monitored traits include typing habit and hand morphology (Feng et al. 2013), graphic touch gesture feature (Zhao et al. 2013), and cardiac motion and geometric information (Lin et al. 2017). Multifactor authentication is another alternative approach to enhance the security of the single trait authentication, which offers multifactor protections in case one credential has been compromised. For example, a set of behavioral biometrics of micro-movements and orientation patterns during hand movement, orientation, and grasp has been proposed for mobile authentication (Sitová et al. 2016). Controls such as one-time passwords, grid-based authentication, and digital-certificate-based authentication schemes can also help augment existing security solutions.

- **Malware detection**

To protect our mobile devices, it is essential to detect the mobile malware and illegal activities including anomaly, misuse, or specification-based system. Basically, there are two kinds of detection techniques as static analysis and dynamic analysis. In static analysis, malicious codes are detected by unpacking and decompiling the application. "DroidMOSS" developed by Zhou et al. (2012) could generate fingerprint for an application and perform similarity test for two "same" apps; then DroidMOSS made sure the application is not affected or repacked with malware to avoid application-based threats on mobile devices, while the dynamic analysis identified the malicious behaviors by running the application on an emulator or a device.

For example, monitoring power consumption is an easy way to detect whether malware are on your smartphone. Since the detection and analysis tasks utilize heavy resources of mobile devices, some of these tasks can be moved to the cloud for processing in the future.

- **Application development guidelines**

The awareness of secure programming guidelines is vital to the secure application development by preventing the occurrence of common errors in programming. Some useful guidelines include performing secure logging and error handling; following the principle of least privilege; validating input data; implementing secure data storage; and avoiding insecure mobile OS features, etc. (Jain and Shanbhag 2012).

- **Hardware-associated protection**

Hardware-associated protection implements a trusted execution environment by designing a hardware-enforced isolated execution environment for security-critical code, which provides protective mechanism in storing and processing sensitive data (Zhang et al. 2016). Techniques such as secure and authenticated boot (Ekberg et al. 2013) are employed to maintain a root of trust on the device. The most popular approach for mobiles is TrustOTP (trust one-time passwords) (Sun et al. 2015) that builds upon the ARM's TrustZone (Brasser et al. 2016) technology. Because it is a hardware-based protection, TrustOTP can prevent attacks that are in the mobile OS and even works when the mobile OS crashes.

References

- Aviv AJ, Gibson KL, Mossop E, Blaze M, Smith JM (2010) Smudge attacks on smartphone touch screens. *Woot* 10:1–7
- Beyah R, Venkataraman A (2011) Rogue-access-point detection: challenges, solutions, and future directions. *IEEE Secur Priv* 9(5):56–61

- Bradley T (2011) Droiddream becomes android market nightmare. PCWorld
- Brasser F, Kim D, Liebchen C, Ganapathy V, Iftode L, Sadeghi AR (2016) Regulating arm trustzone devices in restricted spaces. In: Proceedings of the 14th annual international conference on mobile systems, applications, and services. ACM, pp 413–425
- CCC (2014) Fingerprint biometrics hacked again. <http://www.ccc.de/en/updates/2014/ursel>. Accessed by 13 May 2017
- Celiktutan O, Avcibas I, Sankur B (2007) Blind identification of cellular phone cameras. In: Security, steganography, and watermarking of multimedia contents IX, international society for optics and photonics, vol 6505, p 65051H
- Coursen S (2007) The future of mobile malware. *Netw Secur* 2007(8):7–11
- Damopoulos D, Kambourakis G, Gritzalis S (2011) iSAM: an iPhone stealth airborne malware. In: IFIP international information security conference. Springer, pp 17–28
- Eklberg JE, Kostiainen K, Asokan N (2013) Trusted execution environments on mobile devices. In: Proceedings of the 2013 ACM SIGSAC conference on computer & communications security. ACM, pp 1497–1498
- Feng T, Zhao X, Carbutar B, Shi W (2013) Continuous mobile authentication using virtual key typing biometrics. In: 2013 12th IEEE international conference on trust, security and privacy in computing and communications (TrustCom). IEEE, pp 1547–1552
- Jain AK, Shanbhag D (2012) Addressing security and privacy risks in mobile applications. *IT Prof* 14(5): 28–33
- Johnston D, Walker J (2004) Overview of IEEE 802.16 security. *IEEE Secur Priv* 2(3):40–48
- La Polla M, Martinelli F, Sgandorra D (2013) A survey on security for mobile devices. *IEEE Commun Surv Tutor* 15(1):446–471
- Lin F, Song C, Zhuang Y, Xu W, Li C, Ren K (2017) Cardiac scan: a non-contact and continuous heart-based user authentication system. In: Proceedings of the 23rd annual international conference on mobile computing and networking (MobiCom 17), Snowbird, pp 315–328
- Manabe H, Yamakawa Y, Sasamoto T, Sasaki R (2009) Security evaluation of biometrics authentications for cellular phones. In: Fifth international conference on intelligent information hiding and multimedia signal processing, 2009 (IIH-MSP'09). IEEE, pp 34–39
- Notare MA, da Silva Cruz FA, Riso BG, Westphall CB (1999) Wireless communications: security management against cloned cellular phones. In: Wireless communications and networking conference, 1999 (WCNC 1999), vol 3. IEEE, pp 1412–1416
- Racic R, Ma D, Chen H (2006) Exploiting mms vulnerabilities to stealthily exhaust mobile phone's battery. In: *SecureComm*, vol 6. Citeseer, pp 1–10
- Schlegel R, Zhang K, Zhou Xy, Intwala M, Kapadia A, Wang X (2011) Soundcomber: a stealthy and context-aware sound trojan for smartphones. In: NDSS, vol 11, pp 17–33
- Singh R, Bhargava P, Kain S (2007) Cell phone cloning: a perspective on GSM security. *Ubiquity* 2007:1
- Sitová Z, Šeděnka J, Yang Q, Peng G, Zhou G, Gasti P, Balagani KS (2016) Hmog: new behavioral biometric features for continuous authentication of smartphone users. *IEEE Trans Inf Forensics Secur* 11(5): 877–892
- Statista (2018) Number of apps available in leading app stores as of 1st quarter 2018. <https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/>
- Sujithra M, Padmavathi G (2012) Mobile device security: a survey on mobile device threats, vulnerabilities and their defensive mechanism. *Int J Comput Appl* 56(14): 24–29
- Sun H, Sun K, Wang Y, Jing J (2015) Trustotp: transforming smartphones into secure one-time password tokens. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. ACM, pp 976–988
- Whitney L (2010) Apple sued over privacy in iPhone, iPad apps. <https://www.cnet.com/news/apple-sued-over-privacy-in-iphone-ipad-apps/>
- Xu N, Zhang F, Luo Y, Jia W, Xuan D, Teng J (2009) Stealthy video capturer: a new video-based spyware in 3G smartphones. In: Proceedings of the second ACM conference on wireless network security. ACM, pp 69–78
- Zhang L, Zhu D, Yang Z, Sun L, Yang M (2016) A survey of privacy protection techniques for mobile devices. *J Commun Inf Netw* 1(4):86–92
- Zhao X, Feng T, Shi W (2013) Continuous mobile authentication using a novel graphic touch gesture feature. In: 2013 IEEE sixth international conference on biometrics: theory, applications and systems (BTAS). IEEE, pp 1–6
- Zhou W, Zhou Y, Jiang X, Ning P (2012) Detecting repackaged smartphone applications in third-party android marketplaces. In: Proceedings of the second ACM conference on data and application security and privacy. ACM, pp 317–326

Mobile Social Network

► Data-Driven Mobile Social Networks

Mobility

► Delay-Tolerant Network Routing

Mobility Management

► Mobility Management in NDN

Mobility Management in 5G

Dongmyoung Kim

AIX Center, SK Telecom, Jung-gu, Seoul, Korea

Synonyms

Location management; Reachability management

Definitions

Mobility management in cellular networks is a set of functions that the network uses to keep track of the location and status of the mobile device and to provide proper service to the mobile users.

Historical Background

Mobility management in cellular networks is a set of functions that the network uses to keep track of the location and status of the mobile device and to provide proper service to the mobile users. In the cellular networks, mobile user equipment (UE) may move around whole network coverage, and it is expected that both Mobile Terminated (MT) and Mobile Oriented (MO) services are provided regardless of UE location and movement. Therefore, mobility management is one of the major functions in the cellular networks. Mobility management in cellular networks generally includes user equipment (UE) registration to the network, UE location tracking, paging to enable mobile terminated communication, support of handover, and so on. Additionally, in a wide sense, mobility management also includes the functions to support session continuity during

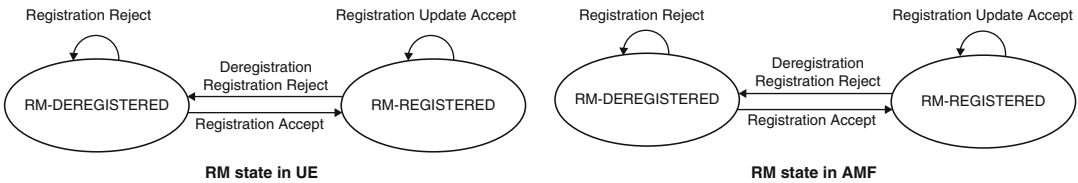
user mobility, e.g., IP address allocation, traffic anchoring, tunneling, and so on. The earlier generations of 3GPP cellular networks, including 3GPP LTE system, have supported the mobility management function (3GPP 2017a). 3GPP 5G network also supports the mobility management functions similar to the ones supported in 3GPP LTE system with some enhancements as specified in 3GPP (2017b,c,d,e). This article summarizes the mobility management in 3GPP 5G network.

Key Mobility Management Functions in 5G Network

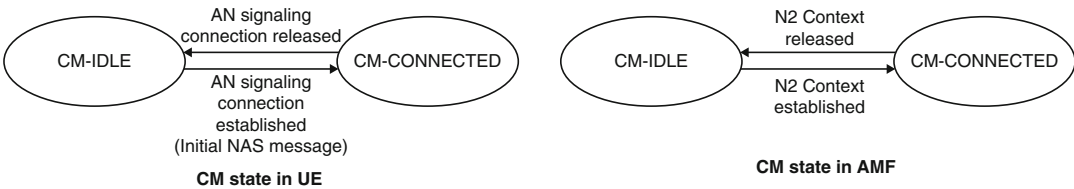
User State Management

In the cellular network, which mobility management functions and procedures are applied to a mobile user depend on state of the user. Three types of states are defined in the 5G system, i.e., Registration Management (RM) state, Connection Management (CM) state, and Radio Resource Control (RRC) state. RM and CM states are managed by the core network. The RM state represents whether a UE is registered in the 5G Core Network (CN), and the CM state shows whether Non-Access Stratum (NAS) signaling connection between UE and 5G CN is established. On the other hand, RRC state is managed by Radio Access Network (RAN), and it represents whether a connection between UE and 5G RAN exists or not. A UE in the RRC-CONNECTED state is in CM-CONNECTED state, and a UE in the RRC-IDLE state is in CM-IDLE state except during the transient phase, because radio link connection is required to establish NAS signaling connection to the core network (Figs. 1 and 2).

RRC-INACTIVE state is newly introduced in the RRC state model. The new state is proposed to be used as a primary sleeping state prior to RRC-IDLE state. When a UE moves to the new state, both the UE and RAN keep the context information of the UE's RRC connection, such as UE capabilities and security context, though the active radio link between the UE and RAN is released. The new state enables a lightweight transition from inactive to active



Mobility Management in 5G, Fig. 1 RM state model in 5G system



Mobility Management in 5G, Fig. 2 CM state model in 5G system

data transmission. A UE in the RRC-INACTIVE state is in CM-CONNECTED state except during transient phase. The 5G CN is not aware of the UE transitions between CM-CONNECTED with RRC-CONNECTED and CM-CONNECTED with RRC-INACTIVE state.

Handover and Cell Reselection

The choice of mobility procedure when the user moves to the coverage of other cell depends on RRC state of the user. The mobile UE in RRC-CONNECTED state uses a network-triggered procedure called handover for cell change, while mobile users in RRC-IDLE state or RRC-INACTIVE state uses a mobile-triggered procedure called cell reselection for cell change.

For the UE in RRC-CONNECTED state, Radio Access Network monitors the signal strength/quality from multiple cells to a user based on measurement report from the user and decides to trigger handover to serve the mobile user in a better cell. RAN and core network is aware of the UE mobility event and UE location after handover.

In the RRC-IDLE state or RRC-INACTIVE state where the user does not maintain active connection with RAN, the UE decides whether to camp on the current cell or to reselect a neigh-

boring target cell based on signal strength measurements without interaction with the network. Therefore, the network may not be aware of the mobility event and exact location of the UE after cell reselection. Therefore, specific UE location tracking mechanisms are needed for the UEs in RRC-IDLE state or RRC-INACTIVE state, and they are described in the next section.

UE Location Tracking and Paging

UE location tracking is responsible for detecting whether the UE is reachable and providing UE location for the network to reach the UE. In most cellular networks, the location update procedure, which allows a mobile device to inform the cellular network when it moves from one location area to the next area, is used as the basic mechanism for location tracking of idle UEs, e.g., Tracking Area Update procedure in 3GPP LTE network. 5G network also supports such location update procedures.

In the 5G network, such functions can be located either at 5G CN (in case of RRC-IDLE state) or 5G RAN (in case of RRC-INACTIVE state). When a UE registers with the network, the network allocates a set of Tracking Areas (TAs) as a registration area of the UE. Each TA is composed of one or multiple cells, and each cell broadcasts the TA which the cell is

contained in. The UE in RRC-IDLE state moving to a new cell checks whether the new cell is still contained in the current registration area based on the broadcasted information. The idle UE triggers a registration update procedure only if the new cell is not contained in the current registration area. Accordingly, 5G (core) network keeps track of the location of UE in RRC-IDLE state in a registration area level.

In 5G system, RAN also needs to support the location tracking for the UE in RRC-INACTIVE state. In that state, RAN needs a new location tracking functionality to detect the location of the UE because the connection between the UE and the RAN is not active. RAN allocates RAN notification area, which is composed of multiple cells, to a UE, and the UE in RRC-INACTIVE state triggers a RAN notification area update procedure only if the new cell is not contained in the current RAN notification area. 5G (radio access) network keeps track of the location of UE in RRC-INACTIVE state in a RAN notification area level.

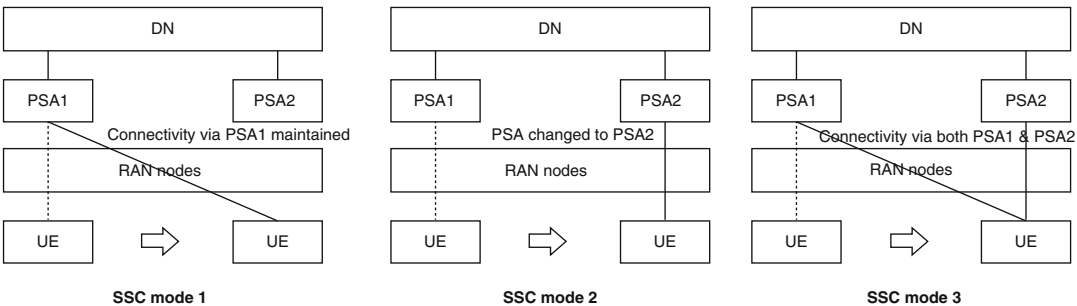
When downlink traffic arrives and the UE is in RRC-IDLE state or in RRC-INACTIVE state, the network (CN or RAN) triggers paging procedures with the consideration of the UE's expected location obtained by the UE location tracking procedures. Then, the UE which detected the paging message sends response to the network, and network gets to know the exact location of the UE, i.e., the current serving cell. The network provides the mobile terminated service to the detected UE location.

Dynamic Mobility Policy Enforcement

Policy Control and Charging (PCC) framework defined in the earlier generation of 3GPP system enables dynamic policy enforcement. In PCC framework of 3GPP LTE system, only session and QoS-related policies are supported. On the other hand, 5G PCC framework also supports mobility policy enforcement. Policy Control Function (PCF), which is the network function mainly managing the policy, can direct interact with AMF, which is the network function managing UE mobility, to support mobility-related policy enforcement. The mobility policy includes Service Area Restrictions and RAT/Frequency Selection Priority.

Support of Session and Service Continuity

Providing session and service continuity of an active session for mobile user is one of the most important tasks of cellular networks. To support session continuity, 5G system provides session anchoring in the gateway and tunneling (tunnel generation and update to the new RAN node on UE mobility) between the anchoring gateway and RAN node like the earlier generation of cellular networks. In 5G system the anchoring gateway (such as P-GW in 4G) is referred to as PDU session anchor (PSA) User Plane Function (UPF). The PSA terminates the user plane in the 5G core network and interfaces with the data network. In case of IP type of session, IP address preservation is an important aspect of session and service continuity, so that the PDU session anchor



Mobility Management in 5G, Fig. 3 SSC mode in 5G system

is responsible for IP anchoring as P-GW does in the EPC.

The main enhancement on session and service continuity in 5G system is that the different levels of session and service continuity procedures can be used according to the characteristics of each session. In the LTE system, continuity of IP session for all UEs is guaranteed in the whole system area. That is, the P-GW and the IP address of UE's PDU session are maintained regardless of the location of the UE such that the session continuity is maintained. On the other hand, 5G system aims to provide various levels of session continuity depending on the type of UE and type of service by allocating different Session and Service Continuity mode (SSC mode) per PDU session.

Three types of SSC modes are defined in 3GPP 5G system, namely, SSC mode 1, SSC mode 2, and SSC mode 3. Separate SSC modes are closely related to how the PDU session anchor is allocated and managed. In SSC mode 1, session continuity is guaranteed in all areas by keeping the PSA identical regardless of access network type and UE location as in the LTE system. In SSC mode 2, the same PSA is maintained across only a subset of the whole network. The PSA of a PDU session may be released and reallocated based on some criteria, e.g., to be served by the better UPF considering the UE's new point of attachment to the network. Lastly, SSC mode 3 enables the UE to connect to a new PSA before the connection with the existing PSA is released, and hence, the UE can receive services from two PSAs at the same time. By using this mode, various optimization can be supported, e.g., a newly created flow is served by a new PSA located in a preferred location while maintaining the existing flows in the previous PSA to provide session continuity (Fig. 3).

Cross-References

- [5G Wireless](#)
- [Future Wireless Network Architecture and Network Slicing](#)
- [Handoffs, Modeling in IP Networks](#)
- [Network-Layer Mobility Management](#)

Acknowledgments This work was supported by the ICT R&D program of MSIP/IITP, [R7116-16-1001, "Development of 5G Core Network Technologies Standards"]

References

- 3GPP (2017a) 3GPP TS 23.401, general packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access, Rel.15
- 3GPP (2017b) 3GPP TS 23.501, system architecture for the 5G system; stage 2, Rel.15
- 3GPP (2017c) 3GPP TS 23.502, procedures for the 5G system; stage 2, Rel.15
- 3GPP (2017d) 3GPP TS 38.300, NR; NR and NG-RAN overall description; stage 2, Rel.15
- 3GPP (2017e) 3GPP TS 38.331, NR; NR and NG-RAN overall description; stage 2, Rel.15

Mobility Management in LTE-U/LAA

- [Handover in LTE-U/LAA](#)

Mobility Management in NDN

Zhiwei Yan
CNNIC, Beijing, China

Synonyms

[Mobility management; NDN](#)

Definition

Named Data Networking, as a clean-slate network model, is designed to support direct routing with content name in order to improve the security and efficiency for the future Internet. In NDN (Zhang et al. 2010), content such as a movie is divided into a set of individually named smaller content objects. Content object names are hierarchical and human-readable similar to the

Routing-Based Approach: In the routing-based approach (Han et al. 2014), two faces corresponding to before and after handover exist in each entry of the FIB. When a router receives an Interest packet, the mobility face (corresponding to the state after handover) of the FIB is firstly checked. After the MP moves, it sends an informing control message to its AR-H. Upon receiving the informing control message, the AR-H sends back an updating control message to update the FIBs along the path from the AR-H to the MP. In this way, the Interest packet can be forwarded to the current MP.

Locator/Identifier Split Approach: In the locator/identifier split approach (Hermans et al. 2012), a new field called Location Name is added to the Interest packets. When an NDN router receives an Interest packet, the Location Name is firstly searched in the FIB. After the MP moves, it sends the binding information to the AR-H, and then the binding information with the MP prefix and its AR-H prefix are recorded in the AR-H. When the AR-H receives an Interest packet that requires the MP's content, it puts the AR-F prefix in the Location Name of the Interest packet. In this way, the Interest packet with the Location Name is forwarded to the MP. The content can be cached with its original content name in the routers.

As stated above, the content name-based routing has no relationship with its location in NDN. Then the mobile consumer can reissue the Interests from the new location to support its mobility. However, how to use this "loose" routing scheme and on-path caching for efficient mobility management has a huge research space (Kim et al. 2012). Do-hyung Kim differentiated the real-time service and unreal-time service of NDN. Then a rendezvous point is deployed for the necessary location management for the mobile consumer (Lee et al. 2011). J. Lee and some other researchers proposed the proxy-based mobility management schemes for mobile consumer in NDN (Lee and Kim 2011; Oh et al. 2010; Wang et al. 2013), mainly to reduce the packet loss during the handover. The main operation can be illustrated as (1) after the consumer detects its

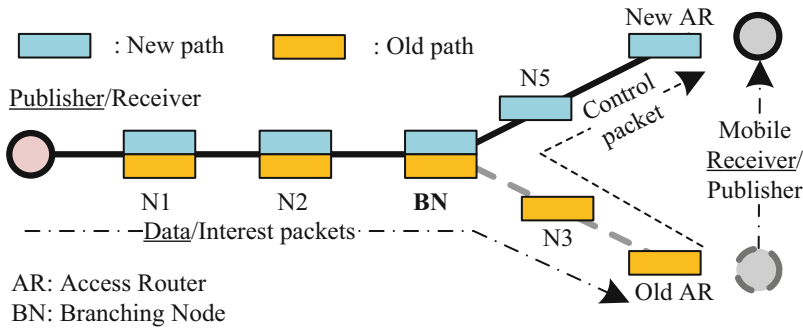
movement, it immediately sends "Hold request" message to its proxy. (2) When receiving the "Hold request" message, the proxy stops delivering data packets and only stores the data packets in its local repository for subsequent retransmissions. (3) When the mobile consumer acquires new location information, it notifies the location information to its proxy node with "Handover notification" message. Then the proxy node transmits the stored data packets to the new location of the mobile consumer.

However, in reality, it's normally required to support mobility management for both consumer and publisher with the same logic because a node may act as consumer and publisher simultaneously. Z. W. Yan proposed a distributed mobility management solution (Yan et al. 2016), which can well support both mobile consumer and mobile publisher and even their simultaneous mobility as shown in Fig. 2.

In this solution, a new singling message named as "Control" message is designed in order to probe the branching node on the paths before and after handover, and it contains necessary information to indicate the new access router and the pending content. Then the on-path routers from the branching node to the new access router will adjust the routing information (PIT for mobile consumer and FIB for mobile publisher) accordingly. In this way, the mobile consumer can get the data as soon as possible after the handover. Besides, the established states can be well used during the handover and then the protocol cost can be reduced. Similarly, the publisher mobility is also based on the branching node probing and hop-by-hop state update, and then the signaling cost caused by the routing flooding can be reduced significantly.

Key Applications

With the development of mobile Internet, more and more traditional networking services will be deployed in the mobile environment. For NDN, the mobile applications supported by the above-mentioned mobility management schemes may be deployed in everywhere, for example, smart



Mobility Management in NDN, Fig. 2 Distributed mobility management in NDN

transport, data dissemination in partially connected environments, collaborative sensing, and presence-aware services (Siris et al. 2012).

Future Directions

With the development of wireless communication technologies, such as 5G, more and more content will be produced and consumed in the high-speed wireless networks. NDN provides a solution to support efficient routing which totally de-couples the content identifier and its location. Then this distributed communication model asks for distributed mobility management scheme (Chan et al. 2014) to guarantee its efficiency and scalability.

Cross-References

- [Distributed IP Mobility Management](#)
- [Handoffs, Modeling in IP Networks](#)
- [Mobile IP](#)
- [Mobility Management with ID/Locator Separation](#)
- [Proxy Mobile IPv6](#)

References

- Chan H, Liu D, Seite P, Yokota H, Korhonen J (2014) Requirements of distributed mobility management. RFC 7333
- Han D, Lee M, Cho K, Kwon T, Choi Y (2014) Publisher mobility support in content centric networks. In: Proceedings of conference on information networking, Phuket, pp 214–219
- Hermans F, Ngai E, Gunningberg P (2012) Global source mobility in the content-centric networking architecture. In: Proceedings of the 1st ACM workshop on emerging name-oriented mobile networking design-architecture, algorithms, and applications, South Carolina, USA
- Johnson D, Perkins C, Arkko J (2011) IP mobility support for IPv6. RFC 6275
- Kim D et al (2012) Mobility support in content centric networks. In: Proceedings of the ICN workshop on Information-centric networking, Helsinki, Finland
- Lee J, Kim D (2011) Proxy-assisted content sharing using content centric networking (CCN) for resource-limited mobile consumer devices. *IEEE Trans Consum Electron* 57(2):477–483
- Lee J, Kim D, Jang MW, Lee BJ (2011) Proxy-based mobility management scheme in mobile content centric networking (CCN) environments. In: Proceedings of the 29th International Conference on Consumer Electronics (ICCE), Las Vegas, USA
- Lee J, Cho S, Kim D (2012) Device mobility management in content-centric networking. *IEEE Commun Mag* 50(12):28–34
- Oh SY, Lau D, Gerla M (2010) Content centric networking in tactical and emergency MANETs. In: Proceedings of the 3rd IFIP, Venice, Italy
- Siris V et al (2012) Content-centric architectures for moving objects. COST Action IC0906 WiNeMO white paper
- Wang L, Waltari O, Kangasharju J (2013) MobiCCN: mobility support with greedy routing in content-centric networks. *Proceedings of IEEE Globecom*
- Yan Z, Zeadally S, Zhang S, Guo R, Park Y (2016) Distributed mobility management in named data networking. *Wirel Commun Mob Comput* 16(13):1773–1783
- Zhang L, Estrin D, Burke J, Jacobson V, Thornton JD, Smetters DK, Zhang B, Tsodik G, Claffy KC, Krioukov D, Massey D, Papadopoulos C, Abdelzaher T, Wang L, Crowley P, Yeh E (2010) Named data networking (NDN) project. NDN Technical Report NDN-0001

Mobility Management with ID/Locator Separation

Ved P. Kafle

Network System Research Institute, National Institute of Information and Communications Technology, Koganei, Tokyo, Japan

Synonyms

ID/locator separation-based mobility management; ID/locator split-based mobility management; Mobility management with ID/locator split

Definition

Mobility management includes network functions that make mobile devices remain connected to the network despite their movement from one place to another, and making them reachable irrespective of their location. Mobility management includes two types of functions: location update and handover. Location update function maintains reachability of mobile devices irrespective of their location, and handover function maintains connectivity during the time when mobile devices detach from one network and attach to another.

ID/locator separation is a concept of using two distinct values (i.e., identifier and locator) for the representation of identity and location of a mobile device. IDs are used in the application and transport layers to identify the communication endpoints or session states, while locators are used in the network layer to denote the location of mobile devices by the routing system in the network topology. This concept is different from the original concept of the Internet, where an IP address is used as both ID and locator. Mobility management with ID/locator separation becomes easier than in conventional IP networks because the change of locators during mobility does not require changing IDs being used in the transport session states.

Historical Background

Mobility was not considered when designing the Internet about 40 years ago. It was assumed that the IP address of a device, host, or node (Note: device, host, and node are used interchangeably in this article) remained static during a communication session. The IP address has been used in the role of both identifiers and locators. Namely, the IP address is used in the network layer protocols as a locator to locate the device in the routing system and forward packets towards it. The same IP address is also used in the transport and application layers to identify the communication sessions or endpoints. Actually, this dual semantic of IP address is hindering the Internet from supporting mobility natively because when the mobile device moves, it has to change its IP address, resulting in terminating the session identified by the IP address.

To enable a mobile device continue its communication session even after moving from one network to another in the Internet, Mobile IP protocols have been developed and standardized by the Internet Engineering Task Force (IETF) in 2002 (for IP version 4, or IPv4) and in 2004 (for IP version 6, or IPv6). Mobile IP protocols allow mobile devices to possess two types of addresses: home address and care-of address. Home address is a persistent address obtained from the home network prefix and used in application and transport layers, whereas care-of address is obtained from a visiting or foreign network and used temporarily in the network layer as long as the mobile device resides in the visiting network. Mobile IP protocols are based on the concept of “map and encapsulate,” where the IP address, i.e., home address, existing in the IP header of packets destined to the home network is mapped with the care-of address by searching in a mapping table, and IP packets are encapsulated by another IP header containing the care-of address in the destination address field and forwarded to the mobile device’s current location.

Mobile IP protocols are complicated due to the necessity of maintaining home agents for managing up-to-date mapping between home addresses and care-of addresses and tunneling

packets. Mobility management with ID/locator separation can be simpler and more efficient because it allows a single ID to dynamically associate with various locators at the same time or different instances of time. It has the potential to natively support heterogeneous types of network-layer protocols and multihoming. These features enable to perform make-before-break or seamless handover easily without requiring any network-side anchor points and tunnel management.

Key Proposals

ID/locator separation-based network technologies and standards have recently been developed by IETF, which has published several RFCs on Host Identity Protocol (HIP), Locator ID Separation Protocol (LISP), and Identifier-Locator Network Protocol (ILNP). Similarly, research projects in various countries have also proposed ID/locator separation-based networking technologies, such as Heterogeneity Inclusion and Mobility Adaptation through Location ID Separation (HIMALIS) in Japan, Mobile Oriented Future Internet (MOFI) in Korea, and MobilityFirst in the USA. Since these proposals are based on the same concept of ID/locator separation, they have many common features, which are discussed next.

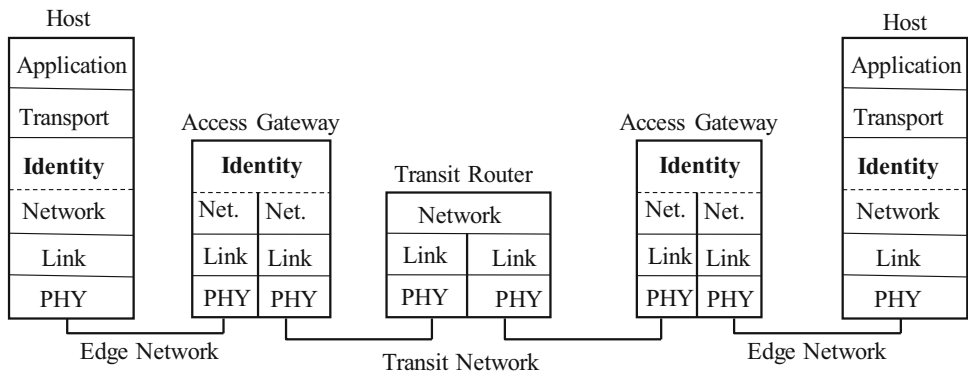
ID/locator separation-based architectures use nontopological values as the identifiers of mobile hosts. They use topologically significant values as locators to represent the location of mobile hosts in the network. The set of locator values associated with a host change when the host's connectivity in the Internet topology changes due to mobility; however, the identifier does not change. Dynamic bindings exist between identifiers and locators values, which are updated as locators change. The bindings are maintained inside the host as well as stored in a mapping service system such as the Domain Name System (DNS).

Figure 1 depicts a union of protocol stacks of all ID/locator split architectures mentioned above. This figure shows the existence of a new layer, which is often known as identity layer, between the transport and network layers in the

host and access gateway protocol stacks. The identity layer isolates the applications and transport layers from the underlying network layer and performs ID-to-locator mapping. Some proposals (such as ILNP and MOFI) do not explicitly mention about the existence of an independent identity layer, but assume similar functions existing in the network layer. In host-based ID/locator separation architectures (such as HIP), the identity layer exists only in the host protocol stack and in the network-based architectures (such as LISP), the identity layer functions exist in the access gateways only, while in the hybrid architectures (such as HIMALIS and MOFI) these functions do exist in both hosts and access gateways.

In ID/locator separation architectures, the application and transport layer protocols use IDs to represent communication endpoints and session states. IDs have end-to-end significance and they do not change due to mobility. On the other hand, locators are used in the network layer for routing and forwarding of packets. The locator values appear in the network header (i.e., IP header) of packets and may have only local significance and change as the packets pass through the access gateways. The set of locator values associated with a node also change when the node's connectivity in the Internet topology changes due to mobility; however, the identifier does not change, thus the transport and transport layers session states remain unchanged after mobility.

Mobility management with ID/locator separation requires involvement of the identity layer, the network layer, as well as the lower layers, when a mobile device moves from one network to another. The lower layers perform the lower-layer handoff functions such as detecting that the radio signal coming from a new network is getting stronger than that from the currently connected network and setting up radio channels and parameters for a new connection. Similarly, the network layer configures a new locator based on the addressing scheme used in the new network and notifies the identity layer about the new locator. The identity layer performs the location update signaling function to update the ID-to-locator mapping cached in the correspondent host



Mobility Management with ID/Locator Separation, Fig. 1 ID/locator protocol stack generic representation

and stored in the mapping server system. After the location update, ongoing communication sessions of the mobile host continue from the new network. Moreover, since ID/locator separation architectures naturally support multihoming by associating the same ID with multiple locators, the mobile host can momentarily be multihomed during the handover time by setting a new link and locator in the new network while still holding the link and locator from the old network. Seamless handover is easily achieved by having the mobile host perform location update in the correspondent host from the new network while still being connected with the old network.

These ID/locator separation-based architectures have some subtle differences in terms of the ID type and configuration methods, ID-to-locator mapping database system, and components handling mobility management functions. These are discussed next.

HIP

HIP architecture and protocol have been specified in RFCs 4423 and 7401, respectively. HIP prescribes to use public keys (and their hash values) as host IDs and IP addresses as locators. HIP extends the Domain Name System (DNS) records to store host IDs as a new resource record type. A host acquires its peer host's ID and locator by sending a domain name resolution request to a DNS server. After obtaining the ID and locator of a mobile host, the correspondent host initializes communication with the mobile host

by exchanging few control packets (known as Base Exchange). Both the source and destination hosts' IDs appear in the identity header and locators in the network header of control packets, while the subsequent data packets include IPsec header instead of the identity header.

For mobility management, HIP introduces rendezvous servers as the anchor points that store the current locators of mobile hosts. Rendezvous server's locators are registered in the DNS server as the mobile host's global locators. The first control packet of an incoming communication request from a correspondent host to the mobile host arrives at the rendezvous server and gets forwarded to the mobile host. The mobile host notifies the correspondent host by sending information about its current locator in the subsequent control packet so that the next packets sent from the correspondent host reach the mobile hosts directly, i.e., without requiring the rendezvous server in the path. On changing its locator due to mobility, the mobile host sends a location update signaling message to update the ID/locator mapping cache of the correspondent host as well as the rendezvous server. In this way, the mobile host can communicate from the new network as well as maintain its reachability by updating the ID/locator mapping stored in the rendezvous server.

ILNP

ILNP architecture has been specified in RFC 6740. It divides 128 bits long IPv6 address into

two 64 bits long parts. The lower 64 bits have been used as identifiers, while the higher 64 bits are used as locators. ILNP assumes that “well-behaved” applications that use FQDN or other nontopological identifiers at the application layer, rather than IP addresses or lower layer identifiers, will perceive no architectural difference between IP and ILNP. Moreover, the ILNP packet format in wire is similar to IP packets; thus, ILNP can operate without requiring ID/locator separation supporting access gateways. However, it assumes the existence of routers that make the routing and forwarding decision on the basis of only the upper 64-bits values of the destination IP address present in the packet header. ILNP also proposes to add new resource record types to DNS server to store the NID (i.e., 64 bits node identifier), L32 (i.e., 32 bits locator), and L64 (i.e., 64 bits locators).

ILNP does not introduce any anchor point for the mobility management. It rather makes the mobile host to perform locator update signaling with the correspondent host as soon as it detects that its locator value has changed. The mobile host is also expected to update its locator values stored in the DNS server. However, since the DNS updates take a longer time to propagate, the mobile host may be unreachable for a while from new correspondent hosts.

LISP

LISP has been specified in RFC 6830. It uses prefix aggregatable endpoint IDs (EIDs), which are also used as locators in the edge network. In the transit network, routing locators (RLOC) are used as locators. EIDs to RLOCs mapping takes place in Ingress and Egress Tunneling Routers (ITR/ETR) located at the border between the edge and transit networks. Both EIDs and RLOCs are syntactically identical to IP addresses; they are different in semantics of how they are used. EID-to-RLOC mapping records are maintained in an overlay network of mapping database, called LISP Alternate Logical Topology (LISP-ALT).

The application and transport layers use EIDs, which also appear in the IP header of packets dispatched from the host. When the packets reach the ITR, the destination EID value present in the

IP header is mapped with a related RLOC obtained from the LISP map server. The IP packets are then encapsulated with LISP header, UDP header, and outer IP header containing the RLOCs in source and destination address fields. When the LISP packets reach at the ETR, their encapsulating outer IP header, UDP header, and LISP header are removed and simply forwarded in the destination edge network by the inner IP header containing EIDs.

Since it also uses EIDs as local locators in edge networks, LISP does not provide host mobility support natively. To provide host-mobility, it proposes to make the LISP mobile host look like a LISP site and possess a lightweight version of the ITR/ETR functions. When the mobile host moves from one network to another, it receives a new RLOC and updates the ITRs that are encapsulating packets from correspondent hosts by using the previous mapping, as well as the map server. The mobile host may use one of the various methods such as by sending a solicit-map-request message or piggybacking mapping data to update the mapping. Nonetheless, LISP lacks the procedure to complete the mapping update process in a short time, thus does not guarantee smooth handover.

HIMALIS

HIMALIS architecture (Kafle and Inoue 2010; Kafle et al. 2014) has the protocol stack exactly as shown in Fig. 1. The identity layer exists in the end hosts as well as in the access gateways that connect the edge network with the transit network. The transit routers are simply IP routers. HIMALIS IDs are 128 bits values containing organizational prefix, scope, and version bits in the prefix.

In HIMALIS, the identity layer obtains mapping between IDs and locators from the mapping system consisting of Domain Name Registry (DNR) and Host Name Registry (HNR) through a name resolution process. That is, these registries are used to resolve hostnames to IDs and locators during a communication initialization phase. The HIMALIS packets contain the identity header along with the transport and network headers. The identity header contains the source

and destination IDs which do not change end-to-end, while the network layer contains the source and destination locators (i.e., IP address) that do change as the packet passes through the access gateway. Having said so, the network header containing the local IP addresses is valid only within the edge network. Thus, a HIMALIS host knows only the local IP addresses of its own and uses the gateway's local IP address as the destination IP address in the network header of outgoing packets. The packets are routed by using local IP addresses in the edge network. As the packets reach the access gateway, their network header is removed and ID present in the identity header is searched in the mapping table to find the corresponding global locators. These global locators are used in the new network header to be attached to the packet. These global locators are used by the core router to route packets in the transit network.

In the destination access gateway, the global IP header is removed and a new local IP header, containing the destination host's and the access gateway's local IP addresses in the destination and source addresses, respectively, is attached to the packet. The packets reach the destination host.

HIMALIS supports mobility across various types of network protocols and differently scoped IP addressing in the edge networks. For example, one edge network can use IPv4 local addressing scheme, while the other edge network can use IPv6 addressing scheme, yet the hosts located in these networks can communicate to each other as the local network protocols are translated by the access gateways. In this way, the application and transport layer functions and session states are completely isolated from the network layer so that changes of network layer protocols and IP address as the host moves from one network to another due to mobility would have no adverse impact on the application sessions if the network layer readdressing is performed instantly during movement. As the host moves to a new network and obtains a new locator from the new access gateway, it immediately performs location update with the correspondent host by sending a signaling message containing its global locator (i.e.,

locator of the new access gateway) so that the correspondent host starts forwarding packets to the new network. At the meantime, the mobile host also informs its previous access gateway about its new locator so that the previous access gateway forwards packets destined to the mobile host to the new access gateway. In this way, HIMALIS attempts to achieve lossless handover.

MOFI

The MOFI architecture (Kim et al. 2013) is designed with three functional features in consideration: global identifier and local locator, query-first data delivery, and distributed control of identifier-locator mapping. MOFI uses global host IDs and local or private IP addresses as locators, and mapping is stored in distributed hash-based ID-locator mapping registers. MOFI does not mention about an explicit identity layer, but divides the network layer into two sublayers: communication and delivery. The communication sublayer is responsible for end-to-end communication, and the delivery sublayer is responsible for routing and forwarding in the access and backbone networks. The communication sublayer performs signaling operation for the locator search from the mapping registers collocated with access gateways.

When the mobile host moves from an old access gateway to a new access gateway, it obtains a new locator. As in HIMALIS, the mobile host informs the new access gateway about the information of the old access gateway so that they establish a handover tunnel to forward packets from the old access gateway to the new access gateway. Then the mobile host sends a location update message to the correspondent host. The mapping registers are also updated with the new locator.

MobilityFirst

The MobilityFirst architecture (Raychaudhuri et al. 2012) focuses on making the network centered on mobility and trustworthiness. It has proposed a globally unique ID (GUID) namespace for network attached objects such as smartphones, people, a group of devices/people, content, or even context. GUIDs are formed from

public keys assigned by a name certification service to the objects. It also introduces a name-based service layer that uses the GUIDs to enable mobility-centric services while ensuring security and trustworthiness. It prescribes a hybrid name/address based routing scheme, employing a fast global name resolution service (GNRS) to dynamically bind the destination GUID to a set of network addresses or locators. Every packet includes GUID in its header so that network nodes can offer GUID-based redirection or late binding to network addresses. On mobility, network addresses do change while GUID remains the same. The change in GUID to address mapping is reflected in GNRS by a mapping update signaling process.

Future Directions

There are several issues related to the deployment of ID/locator separation-based mobility architectures. Among them, security, ID namespace, ID-to-locator mapping directory service, and interoperability with existing IP networks are briefly discussed below.

Security: ID/locator separation-based architectures are considered to enhance security features of mobile networks because they allow mobile devices to use topology-independent names or identifiers to be used in the representation of security context which remain valid despite changing the network layer addresses and locators. HIP, MobilityFirst, and HIMALIS have indicated security as the prominent feature of the new architectures.

ID namespace: Introduction of a network topology-independent IDs and their management is an important issue. Different proposals have considered different types of IDs such as public keys and their hash values (in HIP and MobilityFirst), lower 64-bits of IPv6 addresses (in ILNP), IP address blocks not advertised in BGP routers (in LISP and MOFI), and new IDs (in HIMALIS). A harmonized approach is yet to be initiated.

ID-to-locator mapping directory service: Secured and scalable ID-to-locator mapping

directory service that can provide the mappings in a very low latency while being able to keep the mapping records up-to-date despite frequent changes in locators triggered by mobility is a challenging issue. Different proposals have considered different approaches to maintain the mapping directory service, such as HIP extends DNS records and introduces rendezvous servers, LISP specifies the LISP-ALT, and HIMALIS proposes to use HNRs.

Interoperability with IP network: Due to introduction of new ID handling functions, the ID/locator separation-based mobility management networks need additional proxy functions for interoperability with each other and with the traditional IP networks. For example, there are proxy functions being proposed for HIP, LISP, HIMALIS, and MOFI. The proxy functions can exist either in host or in gateway or both.

Besides providing better mobility management, ID/locator separation-based mobility architectures are also favorable in supporting heterogeneous types of network layer protocols, enhancing security, alleviating BGP routing overhead, and making architecture extendible. These issues have been researched and standardized in various standard development organizations (SDOs) such as IETF and ITU. Although we listed only IETF standards on ID/locator separation networking protocols in the previous sections, there are similar work in ITU (for detail, please see ITU-T Recommendations Y.2015, Y.2057, Y.3032, and Y.3034).

Cross-References

- [Distributed IP Mobility Management](#)
- [Mobile IP](#)
- [Mobility Management in 5G](#)
- [Network-layer Mobility Management](#)

References

- Kafle VP, Inoue M (2010) HIMALIS: heterogeneity inclusion and mobility adaptation through locator id separation in new generation networks. IEICE Trans Commun E93-B(3):478–489

- Kafle VP, Fukushima Y, Harai H (2014) ID/locator split-based distributed mobility management mechanism. *Wirel Pers Commun* 76(4):693–712
- Kim J-I, Jung H, Koh S-J (2013) Mobility Oriented Future Internet (MOFI): architecture design and implementation. *ETRI J* 35(4):666–676
- Raychaudhuri D, Nagaraja K, Venkataramani A (2012) MobilityFirst: a robust and trustworthy mobility-centric architecture for the future internet. *ACM SIG-MOBILE Mobile Comput Commun Rev* 16(3):2–13

Mobility Management with ID/Locator Split

- [Mobility Management with ID/Locator Separation](#)

Mobility Session

- [Proxy Mobile IPv6](#)

Mobility Support in IP Multicast

- [Mobility in Multicast](#)

Mobility in Multicast

Seil Jeon
Information and Communication Engineering,
Sungkyunkwan University, Suwon, South Korea

Synonyms

[Mobility support in IP multicast](#); [Multicast mobility](#)

Definitions

Mobility in multicast means the technical ability to provide mobility support in IP multicast com-

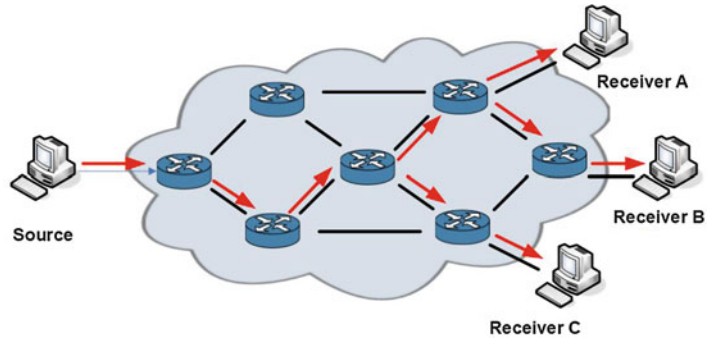
munications. IP multicast session can be resumed after the change of point of attachment by a mobile terminal, without mobility management consideration. But the traditional IP multicasting has two technical issues in a mobility event; one is that the multicast session will be disrupted. Second, a join procedure for the multicast session takes a considerable time for the multicast membership update and multicast routing path update after the IP connection is established. To avoid such time-consuming procedure and session disruption by mobility in IP multicast communication, mobility support mechanisms for IP multicast are required. Depending on mobility management protocol such as Mobile IP or Proxy Mobile IPv6, or other mobility protocols over which IP multicast works, design criteria for mobility management extension are considered.

Historical Background

IP multicast is a well-known transmission technique that efficiently can deliver an IP packet from a sender to multiple receivers by duplicating the packet over IP multicast-enabled networks, without creation and management of multiple IP sessions per content request. Setting up a multicast session is powered by two main technologies: (i) multicast membership group management and (ii) multicast routing. A multicast group is a set of network devices identified by a common multicast address. For the receivers interested in receiving the IP packets of a certain multicast group, it is required to subscribe the membership group using Internet Group Management Protocol (IGMP) for IPv4 (Cain et al. 2002) or Multicast Listener Discovery (MLD) for IPv6 (Vida and Costa 2004). After the interested multicast group information is gathered by the local multicast router, it should send a join message with the multicast group information to the upstream multicast router for establishing a multicast branch between the two multicast routers on the requested multicast group, using multicast routing protocol such as PIM-SM and PIM-DM (Fenner et al. 2016; Adams et al. 2005) or IGMP/MLD Proxy (Fenner et al. 2006) (Fig. 1).

Mobility in Multicast,

Fig. 1 IP multicast transmission to multiple receivers



As wireless/mobile network is evolving and growing with the demand of IP multimedia service, mobility in multicast began to study in earnest with the advent of Mobile IP for the IP session continuity support of a mobile terminal. There are a variety of solutions for mobility in multicast, addressing and/or mitigating the tunnel convergence issue, join latency, point of failure, and so on. In Romdhani et al. (2004), IP multicast support mechanisms over mobile networks were comprehensively investigated, dealing with technical challenges and solutions for mobile source and mobile receiver over Mobile IP (MIP) and its variants such as Fast Mobile IPv6 (FMIPv6) and Hierarchical Mobile IPv6 (HMIPv6).

As Proxy Mobile IPv6 (PMIPv6) addressing the limitations of the host-based mobility protocols (i.e., Mobile IP) appeared (Gundavelli et al. 2008), the design approach for IP multicast support over PMIPv6 has been investigated in IETF Multicast Mobility (MULTIMOB) WG, with the aim of specifying IP multicast support and optimization solutions aligned with PMIPv6 networks (Schmidt et al. 2011; Zuniga et al. 2013) and handling the behavior of IGMPv3/MLD for wireless networks and mobility (Asaeda et al. 2012). All the produced RFCs in IETF MULTIMOB WG are as follows:

- RFC 6224: Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains
- RFC 6636: Tuning the Behavior of the Internet Group Management Protocol (IGMP) and

Multicast Listener Discovery (MLD) for Routers in Mobile and Wireless Networks

- RFC 7028: Multicast Mobility Routing Optimization for Proxy Mobile IPv6
- RFC 7261: Proxy Mobile IPv6 (PMIPv6) Multicast Handover Optimization by the Subscription Information Acquisition through the LMA (SIAL)
- RFC 7287: Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains
- RFC 7411: Multicast Listener Extensions for Mobile IPv6 (MIPv6) and Proxy Mobile IPv6 (PMIPv6) Fast Handovers

For tackling the problems and limitations of the centralized mobility management, i.e., MIP and PMIPv6, Distributed Mobility Management (DMM) has appeared in IETF DMM WG, analyzing the issues of centralized mobility management and extracting the requirements (Chan et al. 2014). In the requirements, IP multicast issues and problems in the centralized mobility management are analyzed, and design considerations for IP multicast over DMM are given.

Key Points

For IP multicast deployment over the mobile networks, it is essentially required to install one of the two IP multicast agents or both into the mobility management entities. The one is an Internet Group Management Protocol (IGMP)/Multicast Listener Discovery (MLD) Forwarding Proxy (in short IGMP/MLD Proxy) (Fenner et al. 2006); the other is a

multicast routing protocol. Both or each of them can be used for providing multicast service over wireless/mobile networks. Though the mobile network can be deployed with both multicast membership protocol and multicast routing protocol, the multicast membership protocol can solely be used and installed in the mobility management entities, and the multicast traffic distribution in the core network can be implemented and deployed by an operator-specific solution. On the other hand, the multicast routing infrastructure can solely be used for multicast packet distribution over the multicast core network, while the subscription of a certain multicast group can be implemented by an operator-specific solution. The IGMP/MLD Proxy provides an available option to associate the mobile network with a multicast infrastructure. It is lightweight compared to the multicast routing protocol and does not require the multicast routing protocol support. Due to the reason, it has been tried to use IGMP/MLD Proxy in PMIPv6 networks (Schmidt et al. 2011; Zuniga et al. 2013). In any cases chosen from the deployment options, mobility management in multicast should consider the following design issues and KPIs.

- *Multicast Traffic Duplication:* IP multicast works per group, not per terminal, while IP mobility management is provided and handled per terminal with a tunneling method. IP multicast aims to provide an efficient transmission of the same packet for multiple terminals. Suppose that there are multiple terminals attached at the same mobility access router and they are supported by the IP tunnels from their home networks in an IP mobility architecture. IP multicast subscription on the same multicast channel for individual mobile terminal will be delivered through each tunnel and multicast traffic will be duplicate. This problem was described in PMIPv6-based networks with a proposed solution to tackle the problem in Jeon et al. (2009), which has been contributed for an optimized IP mobile multicast solution (Zuniga et al. 2013). This

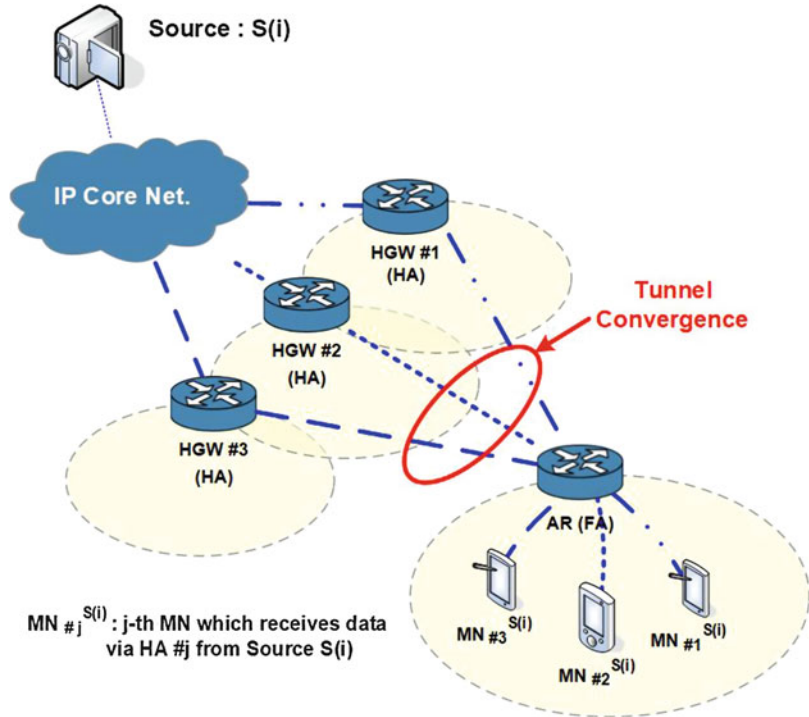
problem has been identified in a distributed mobility environment, as analyzed in Jeon et al. (2012) (Fig. 2).

- *Join Latency:* To receive a copy of IP multicast packet, a receiver should join a group by initiating a multicast membership request when it learns of the group using the membership subscription protocol (IGMP or MLD). The IP multicast router sends a join message to the upstream multicast router to join the group requested from the receiver. Then, an IP multicast branch is established between the two multicast routers. This process takes a considerable time, so reducing the join latency is of importance for seamless service continuity in IP mobile multicast. For addressing the issue, tuning the behavior of IGMP and MLD protocols is required, since the behaviors and parameters of the protocols are originally specified for wired networks (Asaeda et al. 2012). Sending the IP multicast context information to a predicted mobility access router was suggested by extending mobility protocol (Jeon et al. 2009; Figueiredo et al. 2015) or using context transfer protocol (Von Hugo and Asaeda 2013).

Key Applications

- *Mobile IPTV:* Mobile IPTV is a popular application where mobility in multicast can be provided. Nowadays, for receiving multimedia streaming such as drama, movie, and news, IP multicast with mobility can be considered for watching streaming contents, particularly in commute time.
- *Personal Broadcasting Service (PBS):* personal broadcasting service where a mobile sender broadcasts the real-time streaming in a street or moving car is going to be popular. The multicast mobility architecture for PBS was designed and demonstrated in Medieval (2012).
- *Multimedia Broadcast/Multicast Service (MBMS):* MBMS is the 3GPP-based multicast/broadcast platform, which enables multicast/broadcast services for mobile

Mobility in Multicast,
Fig. 2 Multicast traffic
 duplication problem in a
 mobile IP-based network



terminals. MBMS aims for efficiency of multimedia traffic transmission on the same region covered by a base station (BS). However, due to the economies of channel popularity per BS region and dedicated MBMS radio frequency, it is somewhat deprecated.

Cross-References

- [Distributed IP Mobility Management](#)
- [Network-Layer Mobility Management](#)

References

- Adams A, Nicholas J, Siadak W (2005) Protocol independent multicast-dense mode (PIM-DM): protocol specification (Revised). IETF RFC 3973
- Asaeda H, Wu Q, Liu H (2012) Tuning the behavior of the internet group management protocol (IGMP) and multicast listener discovery (MLD) for routers in mobile and wireless networks. IETF RFC 6636
- Cain B, Deering S, Kouvelas I, Fenner B, Thyagarajan A (2002) Internet group management protocol, version 3. IETF RFC 7761
- Chan HA et al. (2014) Requirements for distributed mobility management. IETF RFC 7333
- Fenner B, He H, Haberman B, Sandick H (2006) Internet group management protocol (IGMP)/multicast listener discovery (MLD)-based multicast forwarding ("IGMP/MLD Proxying"). IETF RFC 4605
- Fenner B, Handley M, Holbrook H, Kouvelas I, Parekh R, Zhang Z (2016) Protocol independent multicast-sparse mode (PIM-SM): protocol specification (Revised). IETF RFC 7761
- Figueiredo S, Jeon S, Gomes D, Aguiar RL (2015) D3M: multicast listener mobility support mechanisms over distributed mobility anchoring architectures. J Netw Comput Appl 53:24–38
- Gundavelli S, Leung K, Devarapalli V, Chowdhury K, Patil B (2008) Proxy mobile IPv6. IETF RFC 5213
- Jeon S, Kang N, Kim Y (2009) Mobility management based on proxy mobile IPv6 for multicasting services in home networks. IEEE Trans Consum Electron 55(3):1227–1232
- Jeon S, Figueiredo S, Aguiar RL (2012) A channel-manageable IP multicast support framework for distributed mobility management. In: IFIP wireless days 2012, Dublin
- Medieval F (2012) MEDIEVAL D1.3: final architecture design

- Romdhani I, Kellil M, Lach HY, Bouabdallah A, Bettahar H (2004) IP mobile multicast: challenges and solutions. *IEEE Commun Surv Tutor* 6(1):18–41
- Schmidt T, Waehlich M, Krishnan S (2011) Base deployment for multicast listener support in proxy mobile IPv6 (PMIPv6) domains. IETF RFC 5213
- Vida R, Costa L (2004) Multicast listener discovery version 2 (MLDv2) for IPv6. IETF RFC 3810
- Von Hugo D, Asaeda H (2013) Context transfer protocol extension for multicast. Draft-vonhugo-multimob-cxtp-extension-03
- Zuniga JC, Contreras LM, Bernardos CJ, Jeon S, Kim Y (2013) Multicast mobility routing optimizations for proxy mobile IPv6. IETF RFC 7028

Modeling Approaches for Simulating Molecular Communications

L. Felicetti, M. Femminella, and G. Reali
Department of Engineering, University of Perugia, Perugia, Italy

Synonyms

Molecular communication simulators; Simulation of diffusion-based molecular communications

Definition

Molecular communication involves biological entities that are able to transmit and receive molecules, which represent the information signal.

Historical Background

The molecular communications (MolCom) represent an emerging research area that consists of transmission of information by means of exchange of molecules, carried out by either natural or artificial nanomachines. The physical mechanisms that allow transferring information at such small scales are typically inspired by the biological mechanisms that exist in living

bodies to exchange many types of signaling molecules, such as proteins, pheromones, and immune system activation signals, both within and between different cells. The diffusion-based MolCom are the most studied mechanisms; they are based on the molecule propagation in the fluid medium according to the laws of diffusion (Philibert 2006), without the presence of any flow or drift.

In recent years, a multitude of significant breakthroughs occurred in some strategic socioeconomic fields (Akyildiz et al. 2008). They encompassed multidisciplinary research, bringing together information and communication technologies, molecular biology, physics, chemistry, biotechnology, environmental control, and material science. Together, they will allow realizing the vision of transferring information within biological environments at extremely small scales, down to a size comparable to that of molecules. In this case MolCom are considered an alternative approach to electromagnetic communications due to their unique features of biocompatibility and minimal invasiveness, which are essential for them to be used in living bodies.

Key Applications

Up to now, most interesting potential applications have been identified in the biomedical field (Felicetti et al. 2016). The emerging applications of MolCom techniques focus on both emulations and the development of body area nanonetworks oriented to the diagnosis and treatment of diseases.

As for the first one, the emulation of nanoscale biological processes allows achieving personalized predictions of the evolution of diseases, starting from a limited number of biomarkers, avoiding any traditional in vivo tests on patients (Hood et al. 2011; Bragazzi 2013) helping medical personnel to identify optimal treatments. The computational models are based on a detailed characterization of the molecular communication parameters done by using results of in vitro and in vivo experiments.

For the latter one, the design of nano-sensors/actuators could allow the detection and treatment of a large set of diseases (e.g., cardiovascular diseases, tumors, etc.). In fact, these nano-devices could activate the immune system and/or trigger drug delivery systems in small specific areas without affecting the rest of the body (i.e., smart drugs).

However, other fields, such as food production, functionalized materials and fabrics, as well as environmental and military applications, can be envisioned (Akyildiz et al. 2008).

Communication Models

The most simple and energy-efficient propagation model, without any external cause, is the diffusion-based molecular communication system. In the last years, several researches have been focused on the theoretical analysis and mathematical modeling of the propagation medium, on the receiver modeling (Yilmaz et al. 2014), and on the reception techniques.

In diffusion-based systems, the transmit signal is encoded on the physical characteristics of information molecules (e.g., cytokines, hormones, DNA). Those are able to randomly propagate in the fluid medium via diffusion spreading in the environment according to the negative gradient of the concentration with a velocity that depends on the thermal energy and on the particle size.

The information can be encoded on the concentration of emitted particles, on the frequency of emission, and on the type of released molecules.

The information molecules can interact with the receiver node in several ways, changing or not their concentration around the receiver. The simplest reception model assumes the ideal passive (or transparent) receiver which is permeable to the hitting molecules and is capable only to count the number of the molecules inside its volume. This means that the hitting molecules may unavoidably contribute to the received signal more than once in different symbol intervals

and the receiver may encounter high intersymbol interference (ISI) (Deng et al. 2017).

In a more realistic molecular communication system, each molecule is removed from the environment after the adsorption, and it contributes only once to the received signal. The adsorption takes place through the surface receptors that cover the external membrane of the receiver node. These receptors are able to react only with specific types of information molecules. In more detail, these receptors may adsorb or bind with these molecules. In the first case, the chemical complex formed by the molecule (known also as ligand) and the receptor is internalized by the nanomachine and follows a process known as trafficking (Lauffenburger and Linderman 1996). On the latter case, it may happen that some molecules desorb from their receptors and may contribute several times to the received signal.

In what follows, the main receiver models for a 3D diffusion-based molecular communication system in a fluid environment are analyzed, from the more complex and general to the simplest one. As already mentioned, in the most general case, the information molecules released from the transmitter can be adsorbed and desorb from the surface of the receiver, and the net number of adsorptions is counted for information decoding. This general model will be introduced first, and then the other models will be derived from it.

The overall system consists of a point transmitter and a spherical receiver in an unbounded homogeneous environment. The point transmitter is located at a distance r_0 from the center of the receiver and is at a distance $d = r_0 - r_r$ from the nearest point on the surface of the receiver with radius r_r . The surface of the receiver node is assumed to be covered by a number of compliant receptors that can adsorb at most one molecule at a time. It is assumed that the spherical receiver has no physical limitation on the number or placement of receptors on its surface. Thus, there is no limit on the number of molecules adsorbed to the receiver surface. This is an appropriate assumption for a sufficiently low number of adsorbed molecules, or for a sufficiently high concentration of receptors.

Emission Process

The point transmitter emits the information molecules at $t = 0$. Fick's diffusion equation describes the propagation of the information molecules in the environment:

$$C(r, t \rightarrow 0 | r_0) = \frac{1}{4\pi r_0^2} \delta(r - r_0), \quad (1)$$

where $C(r, t \rightarrow 0 | r_0)$ is the molecule distribution function at time $t \rightarrow 0$ and distance r from the center of the receiver with initial distance r_0 . The first boundary condition is:

$$\lim_{r \rightarrow \infty} C(r, t | r_0) = 0, \quad (2)$$

such that at arbitrary time, the molecule distribution function equals zero when r goes to infinity.

Propagation via Diffusion

The released information molecules randomly diffuse in the medium by means of Brownian motion (Berg 1993) colliding with other molecules. In the case in which the concentration of information molecules could be assumed to be sufficiently low, the collisions between information molecules could be ignored (Berg 1993). This means that each information molecule diffuses independently with constant diffusion coefficient D .

Fick's second law describes the propagation model in a 3D environment (Yilmaz et al. 2014; Philibert 2006):

$$\frac{\partial C(r, t | r_0)}{\partial t} = \nabla^2 (D \cdot C(r, t | r_0)) - k_d C(r, t | r_0). \quad (3)$$

In (3), we have also included the negative contribution of molecule degradation. In fact, signaling molecules released in the environment may degrade with a certain probability and transform into molecules not recognized by the receiver. k_d is the degradation reaction coefficient (time^{-1}), which is the time constant of this chemical phenomenon. In the following, we assume no molecule degradation, i.e. $k_d = 0$.

A more general case considers also the presence of a non-negligible flow in the environment, which often is also constrained in a limited space. This case, that will be not treated in this chapter for space limitation, is well modeled by the advection-diffusion equation (Gentile et al. 2008).

Reception

A generic receiver node should be capable of adsorbing incoming molecules that hit its surface or reflect them back into the fluid environment, based on the adsorption rate k_1 ($\text{length} \times \text{time}^{-1}$), which depends on the chemical affinity of each molecule with the surface receptors of the receiver node. The adsorbed molecules either desorb or remain stationary at the surface of the receiver, based on the desorption rate k_{-1} (time^{-1}).

The adsorption and desorption reactions that can occur on its surface give the second boundary condition of the information molecules:

$$D \frac{\partial (C(r, t | r_0))}{\partial r} \Big|_{r=r_r^+} = k_1 C(r_r, t | r_0) - k_{-1} C_a(t | r_0), \quad (4)$$

where $C_a(t | r_0)$ is the average concentration of molecules that are adsorbed to the receiver surface at time t . The change in the adsorbed concentration over time is equal to the flux of diffusion molecules toward the surface:

$$\frac{\partial C_a(t | r_0)}{\partial t} = D \frac{\partial (C(r, t | r_0))}{\partial r} \Big|_{r=r_r^+}. \quad (5)$$

Combining (4) and (5), the radiation boundary condition is obtained, and it shows that the equivalent adsorption rate is proportional to the surface molecule concentration:

$$\frac{\partial C_a(t | r_0)}{\partial t} = k_1 C(r_r, t | r_0) - k_{-1} C_a(t | r_0). \quad (6)$$

At $t = 0$, there are no information molecules at the receiver surface, so the second initial condition is:

$$\begin{aligned} C(r_r, 0|r_0) &= 0, \\ C_a(0|r_0) &= 0. \end{aligned} \quad (7)$$

Depending on the values assumed by both k_1 and k_{-1} , it is possible to define how the receiver node works, as defined below:

- Both k_1 and k_{-1} are non-zero finite constants: (4) is the boundary condition for the adsorption/desorption receiver. A complete transient and steady-state solution for this quite general case can be found in Deng et al. (2015). A variation of this model, taking into account also the receptor occupancy status, is illustrated in Pierobon and Akyildiz (2011).
- k_1 is a non-zero finite constant and $k_{-1} = 0$, and (4) is the boundary condition for the partial absorbing receiver (or receiver with finite receptors). The partially absorbing receiver with finite receptors fits with the realistic case in which the complex formed by the ligand and the receptor is internalized. Clearly, this may happen only when a molecule binds to one of the compliant surface receptors; otherwise it will be reflected away. A quite complete treatment of the solution of this specific case can be found in Akkaya et al. (2015) and references therein.
- $k_1 \rightarrow \infty$ and $k_{-1} = 0$: (4) is the boundary condition for the fully adsorbing receiver. In this configuration, each colliding molecule with the surface of the receiver will be adsorbed without any further desorption. This configuration models the case in which the surface of the receiver node is largely covered by receptors compliant with ligand molecules. A quite complete treatment of the solution of this model can be found in Yilmaz et al. (2014).
- Finally, a further popular model, which is slightly different from those presented above, is the transparent (or virtual) receiver. In this model, a receiver is able to simply count the molecules crossing its volume, *without any interaction* with them (i.e., adsorption, desorption, or even bounces). Clearly, in this case the coefficients k_1 and k_{-1} are not used, and the

only boundary condition is (2). For this type of receiver, an interesting analysis is presented in Llatser et al. (2013).

The probability of finding a molecule at distance r and time t is given by the time-varying spherically symmetric spatial distribution $C(r, t|r_0)$. The cumulative fraction of adsorbed molecules at time t can be expressed as:

$$F(\Omega_{r_r}, t|r_0) = \int_0^t 4\pi r_r^2 D \frac{\partial C(r, \tau|r_0)}{\partial r} \Big|_{r=r_r} d\tau \quad (8)$$

where Ω_{r_r} represents the surface of the spherical receiver with radius r_r . Upon an emission of Q molecules at time $t_0 = 0$, the average number of adsorbed molecules after t seconds will be $N(\Omega_{r_r}, t|r_0) = QF(\Omega_{r_r}, t|r_0)$. The asymptotic concentration of adsorbed molecules can be found as $t \rightarrow \infty$ for the different formulations of $N(\Omega_{r_r}, t|r_0)$, depending on the values of k_1 and k_{-1} . As for the transparent receiver, the solution is simpler, and the average number of “counted” molecules is given by $N(\Omega_{r_r}, t|r_0) = C(0, t|r_0) V_r$, where V_r is the volume of the spherical transparent receiver.

For a treatment more focused on the chemical species involved into the adsorption and desorption reaction, that is the molecules or ligands, the receptors, and the complex obtained by their binding, the interested reader can refer to Lauffenburger and Linderman (1996).

Mapping Models on Simulation Platforms

BiNS Biological and Nanoscale Communication Simulator

General description BiNS is a simulation package for MolCom systems developed at the University of Perugia (Felicetti et al. 2012).

Its customizable design provides a set of tools which allow creating objects modeling the behavior of biological entities. They can be regarded as either nodes (transmitters, receivers), carriers, or

surrounding objects (e.g., vessel walls). In addition, BiNS permits to configure the properties of the simulated communication channel (e.g., the blood stream or the environment of an in vitro experiment) with the desired accuracy.

The simulator has been implemented in Java and contains generic type of software object, named Nano Object. Nodes and carriers are specific implementations of the Nano Object, and, although they share its general features, they can exhibit very different functions.

The simulation is organized in discrete time steps. Each step consists of a number of phases, in which software objects are triggered in order to execute the operations associated with their specific behavior. The main phases are:

- transmission phase
- reception phase
- information processing phase,
- motion phase
- object destruction phase (during which objects are removed due to lifetime expiration or because they exited from the area of interest)
- collision management phase, which implements the elementary interaction between particles.

BiNS has been validated through wet-lab experiments related to cardiovascular medicine.

Technical requirements and scalability The list of simulated nano-objects is split into smaller lists, which can be handled in parallel by a thread pool.

The simulator uses a fine-grained approach for handling collisions between objects. A collision can produce either a bounce or an assimilation. The latter happens only when a carrier collides with a compliant receptor on the surface of a node. The receptors can implement a non-negligible absorption time; thus a compliant particle could find a receptor in either busy or free state. The simulated environment can be either unbounded or bounded by a surface of custom shape, referred to as simulation domain. For example, in order to simulate communica-

tions within a blood vessel, a cylindrical volume was used. Within a domain, the octree paradigm allows distributing the workload associated with the management of the objects in the domain volume to different threads. There is an ongoing effort for the GPU porting of the collision handling algorithm in order to reduce the computational time significantly.

N3Sim

General description N3Sim (Llatser et al. 2011) is a Java-based complete simulation framework for diffusion-based molecular communications, which allows the evaluation of the communication performance of molecular networks with several transmitters and receivers in an infinite space with a given concentration of molecules. The transmitters encode the information by releasing particles into the medium, thus varying the concentration rate in their vicinity. The diffusion of particles through the medium is modeled as Brownian motion, taking into account particle inertia and collisions among particles. Finally, the receivers decode the information by sensing the local concentration in their neighborhood. It implements a three-layer architecture. The user interface layer interacts with the user to read the input data for the simulation, while the data layer writes the simulation results to files. The domain layer contains the intelligence of the system, i.e., the molecular communication model. N3Sim allows automating the execution of multiple simulations in a simple manner by means of user-defined scripts.

Technical requirements and scalability N3Sim is designed to be executed on a single machine with Java 1.6 or higher. The simulator has been tested both on Linux and Windows. The scalability of the simulator can be improved by selecting a larger value of the simulation time step (at the expense of the accuracy) or by deactivating the collisions among molecules in scenarios with a low molecular concentration. As an example, simulations with up to 10^6 simultaneously emitted molecules have been

successfully performed. The time granularity of a simulation is defined by the user by choosing the simulation time step in the configuration file (typically a few *ms*). The simulation space can be either bounded or unbounded, and both two-dimensional and three-dimensional configurations are supported.

NS-3 Based Simulators

General description NS-3 is a discrete-event network simulator for Internet systems, which was not originally developed for MolCom simulators. Nevertheless, its flexible structure has allowed implementing some basic elements of MolCom. In particular there are two simulation tools that have been developed in NS-3: one in the framework of the IEEE P1906.1 working group (Bush et al. 2015; IEEE Std 1906.1-2015 2016) and another specific implementation for bacteria communications (Jian et al. 2017).

NS-3 is organized in different software libraries that can work together. User programs, written in C++ or Python programming languages, can adapt these libraries or be linked with them. External animators and data analysis and visualization tools are available. Nevertheless, in order to exploit the full potentials of the NS-3, the command line interface is suggested.

Technical requirements and scalability The simulator is designed to be executed on a single machine. The minimal requirements to execute basic simulations are a gcc or clang compiler and Python interpreter. The simulation granularity is defined by users. Event design can be adapted to any granularity and is synchronized in the simulator implementation processing loop.

AcCoRD

General description The Actor-based Communication via Reaction-Diffusion (AcCoRD) is an open-source generic reaction-diffusion MolCom simulator developed in C language (Noel et al. 2017). The actors can be both transmitter and receivers, and the simulation environment is defined as a collection of microscopic and

mesoscopic regions that differ for the simulation accuracy and the computational cost. In the microscopic region, each molecule is individually modeled in each time step, whereas in the mesoscopic one, the total amount of molecules in each region is analyzed. Chemical reactions (e.g., molecule degradation, reversible/irreversible surface binding, etc.) can be defined in the propagation environment.

Technical requirements and scalability

Microscopic and mesoscopic regions can be combined and dimensioned according to the simulation needs, balancing the local accuracy with the computational cost. Also a visualization tool developed in MATLAB is included in the simulator allowing an offline visualization of the simulated environment.

Simulators for Specific MolCom Environments

In addition to the general simulation platforms presented in the section before, there are a number of more specific simulation packages targeted to specific environments. Examples of these simulators include platforms dealing with bacteria colonies (Jian et al. 2017; Wei et al. 2013) introduced also in section “NS-3 Based Simulators”, calcium signaling (Barros et al. 2015), or nervous systems (Malak et al. 2014).

MolCom Simulators Functional Comparison

A comparison of the main simulators that have been developed to characterize the broad set of MolCom systems is sketched in Table 1, where the compatibility with the different receiver models introduced above is illustrated.

However, this is a rough classification of the available simulation tools. In fact, more realistic and refined models can be simulated only by a subset of these ones. For instance, when the interactions between molecules are significant, diffusion equations may become nonlinear (Aranovich and Donohue 2005), regardless of the receiver model. Also, the partial adsorbing receiver can be modeled with a finite adsorption time for each receptor, which is more realistic.

Modeling Approaches for Simulating Molecular Communications, Table 1 Compatibility of the most popular MolCom simulators with the reception models

Simulator	Adsorption/desorption receiver	Partial adsorbing receiver	Full adsorbing receiver	Transparent receiver
BiNS ^a	Easy to upgrade ^c	Yes	Yes	Yes
N3Sim ^b	No	No	Yes	Yes
NS-3 ^c	No	No	No	Yes
AcCoRD ^d	Yes	Yes	Yes	Yes

^a Available on <http://conan.diei.unipg.it/lab/index.php/product/biological-nanoscale-simulator-bins2>

^b Available on <http://www.n3cat.upc.edu/tools/n3sim/download>

^c Available on <https://www.nsnam.org/releases/>

^d Available on <https://warwick.ac.uk/fac/sci/eng/staff/ajgn/software/accord/downloads/>

^e A minor upgrade is necessary to implement desorption by applying the Algorithm 1 illustrated in Deng et al. (2015)

These more complex but also realistic systems can be simulated by BiNS2. At the same time, the desorption functionality is already implemented in AcCoRD, whereas BiNS2 needs an upgrade to implement it.

Cross-References

- [Applications of Molecular Communication Systems](#)
- [Drug Delivery via Nanomachines](#)
- [Molecular Communication for Wireless Body Area Networks](#)
- [Nanonetworks](#)
- [Reaction-Diffusion Channels](#)
- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

References

- Akkaya A et al (2015) Effect of receptor density and size on signal reception in molecular communication via diffusion with an absorbing receiver. *IEEE Commun Lett* 19(2):155–158. <https://doi.org/10.1109/LCOMM.2014.2375214>
- Akyildiz IF et al (2008) Nanonetworks: a new communication paradigm. *Comput Netw* 52(12):2260–2279. <https://doi.org/10.1016/j.comnet.2008.04.001>
- Aranovich GL, Donohue MD (2005) Diffusion equation for interacting particles. *J Phys Chem B* 109(33):16062–16069
- Barros T et al (2015) Comparative end-to-end analysis of ca2+-signaling-based molecular communication in biological tissues. *IEEE Trans Commun* 63(12):5128–5142. <https://doi.org/10.1109/TCOMM.2015.2487349>
- Berg H (1993) *Random walks in biology*. Princeton University Press, Princeton
- Bragazzi NL (2013) From p0 to p6 medicine, a model of highly participatory, narrative, interactive, and “augmented” medicine: some considerations on Salvatore Iaconesi’s clinical story. *Patient Prefer Adherence* 7:353–359. <https://doi.org/10.2147/PPA.S38578>, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3640773/>, ppa-7-353[PII]
- Bush SF et al (2015) Defining communication at the bottom. *IEEE Trans Mol Biol Multi-Scale Commun* 1(1):90–96. <https://doi.org/10.1109/TMBMC.2015.2465513>
- Deng Y et al (2015) Modeling and simulation of molecular communication systems with a reversible adsorption receiver. *IEEE Trans Mol Biol Multi-Scale Commun* 1(4):347–362. <https://doi.org/10.1109/TMBMC.2016.2589239>
- Deng Y et al (2017) Analyzing large-scale multiuser molecular communication via 3-d stochastic geometry. *IEEE Trans Mol Biol Multi-Scale Commun* 3(2):118–133. <https://doi.org/10.1109/TMBMC.2017.2750145>
- Felicetti L et al (2012) A simulation tool for nanoscale biological networks. *Nano Commun Netw* 3(1):2–18
- Felicetti L, Femminella M, Reali G, Lio P (2016) Applications of molecular communications to medicine: a survey. *Nano Commun Netw* 7:27–45
- Gentile F et al (2008) The transport of nanoparticles in blood vessels: the effect of vessel permeability and blood rheology. *Ann Biomed Eng* 36(2):254–61
- Hood L et al (2011) Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nat Rev Clin Oncol* 8:184 EP <https://doi.org/10.1038/nrclinonc.2010.227>, perspective
- IEEE Std 19061-2015 (2016) IEEE recommended practice for nanoscale and molecular communication framework. *IEEE Std 19061-2015*, pp 1–64. <https://doi.org/10.1109/IEEESTD.2016.7378262>
- Jian Y et al (2017) nanoNS3: a network simulator for bacterial nanonetworks based on molecular communication. *Nano Commun Netw* 12:1–11. <https://doi.org/10.1016/j.nancom.2017.01.004>, <http://www.sciencedirect.com/science/article/pii/S1878778916300941>

- Lauffenburger D, Linderman J (1996) Receptors: models for binding, trafficking, and signalling. Oxford University Press, New York
- Llatser I et al (2011) Exploring the physical channel of diffusion-based molecular communication by simulation. In: IEEE GLOBECOM 2011. <https://doi.org/10.1109/GLOCOM.2011.6134028>
- Llatser I et al (2013) Detection techniques for diffusion-based molecular communication. IEEE J Sel Areas Commun 31(12, supplement): 726–734
- Malak D et al (2014) Communication theoretical understanding of intra-body nervous nanonetworks. IEEE Commun Mag 52(4):129–135. <https://doi.org/10.1109/MCOM.2014.6807957>
- Noel A et al (2017) Simulating with accord: actor-based communication via reaction diffusion. Nano Commun Netw 11:44–75. <https://doi.org/10.1016/j.nancom.2017.02.002>, <http://www.sciencedirect.com/science/article/pii/S1878778916300618>
- Philibert J (2006) One and a half century of diffusion: Fick, Einstein, before and beyond. Diffus Fundam 4:6.1–6.19
- Pierobon M, Akyildiz I (2011) Noise analysis in ligand-binding reception for molecular communication in nanonetworks. IEEE Trans Signal Process 59(9):4168–4182
- Wei G et al (2013) Efficient modeling and simulation of bacteria-based nanonetworks with BNSim. IEEE J Sel Areas Commun 31(12):868–878. <https://doi.org/10.1109/JSAC.2013.SUP2.12130019>
- Yilmaz HB et al (2014) Three-dimensional channel characteristics for molecular communications with an absorbing receiver. IEEE Commun Lett 18(6):929–932. <https://doi.org/10.1109/LCOMM.2014.2320917>

Modulation in Molecular Signaling

H. Birkan Yilmaz¹, M. Şükrü Kuran², and İlker Demirkol¹

¹Department of Telematics Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain

²Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey

Synonyms

[Modulation techniques for synthetic molecular communications](#)

Definitions

Molecular signaling is a way of communicating via use of molecules. There are still many applications and environments where the classical technologies are not convenient or appropriate. Inspired by nature, one possible solution to these problems is to use chemical signals as carriers of information, which is called molecular communication (MC). Modulation in *synthetic* MC is the way of embedding the information on one or more properties of molecule emission process.

Historical Background

In 1959, Richard Feynman gave a lecture at American Physical Society meeting in Caltech that is titled “There is Plenty of Room at the Bottom” (Feynman 1959). The talk was pointing out the manipulation and controlling things at a small scale with a huge potential to advance the knowledge frontier. Over the past couple of decades there have been considerable advancements in the fields of nanotechnology, biotechnology, and microrobotics, where designing and engineering of microscale or nanoscale devices begin to take shape. On the other hand, the cooperation and coordination of these tiny devices necessitates the miniaturization of communication networks, in which one of the techniques is MC.

Molecular signaling/communication is at its infancy phase having been first proposed in 2005 from an engineered communication perspective (Nakano et al. 2005). Many theoretical and simulation-based studies are published in the molecular communications (MC) domain. In this entry, the works related to MC modulation techniques are given in chronological order and from simple to complex systems. Readers may feel the chronological order while reading the entry.

Background on Molecular Communication

In MC, molecules are utilized to convey information among communication nodes at micro-

and macroscales (Farsad et al. 2016b). Some of the microscale MC systems can be listed as quorum sensing in bacteria (Cobo and Akyildiz 2010; Abadal and Akyildiz 2011), neurotransmitter signaling in neuromuscular junctions (Kuran et al. 2013), and inter-cell calcium signaling (Nakano et al. 2005; Kuran et al. 2012). Examples of macroscale MC systems in nature include pheromone signaling among plants, chemical signaling among animals, and odor tracking of blue crabs in the ocean (Zimmer and Butman 2000).

In MC, information molecules propagate in a fluid environment via various processes (such as diffusion), and they arrive at the receiver node in a probabilistic manner in which the received molecules constitute the received molecular signal (Farsad et al. 2016b). In its basic form, an MC system mainly consists of a transmitter node, fluid environment, information molecules, and a receiver node (Nakano et al. 2013; Farsad et al. 2016b). In the following subsections, first the common performance metrics for MC channels are given followed by the types and dynamics of the modulation techniques.

Common Metrics for Molecular Communication

One of the first metrics to evaluate the modulation techniques that have been used in MC literature is the system response. This metric mainly focuses on the number of received molecules with respect to time. It is generally used to show the effects of different channel types (e.g., free diffusion, vessel-like environment, constrained diffusion) and transmitted molecular signal waveforms (e.g., sinusoidal, square) on the received signal.

Other common metrics can be listed as bit error rate (BER), symbol error rate (SER), and the channel capacity. In the MC literature, as a performance metric, BER or SER is usually evaluated against several system parameters such as signal-to-noise ratio (SNR), transmitted power, symbol duration, communication distance, and diffusion coefficient. As for SNR, the key chal-

lenge is to clearly define what is the definition of noise in this domain and which noise sources to consider (Nakano et al. 2013). Channel capacity metric is evaluated based on the given error rates and the channel model and provides the achievable data rate of the communication system in question. A key point regarding the correctness of this metric is the selection of the appropriate channel model. Currently, most of the MC literature assume a memoryless MC channel to evaluate the channel capacity. However, as described by Genc et al., due to the diffusion-based dynamics of the MC channel, the memoryless assumption is far from appropriate in the MC domain, and instead a model with memory is required (Genc et al. 2016).

After briefly mentioning these common performance metrics, the rest of the entry will be focused on the types of modulation techniques used in MC systems. At the end of this entry, a table of studies are given in a table with also mentioning the metrics used in each study.

Modulation Techniques

All MC modulation techniques depend on releasing special molecules called messenger molecules (MM) from the transmitter and modulating the bit values of a given message upon the features of these MMs. There are works in the MC literature on the selection of feature(s) to be used to represent the information. These works can be broadly classified into four groups:

Concentration-based Techniques Information is embedded upon the varying concentration level of the transmitted signal.

Type-based Techniques Information is embedded upon the type of the MMs of the transmitted signal.

Timing-based Techniques Information is embedded upon the MM release time instances.

Hybrid Techniques Techniques utilizing more than one of the three features (concentration,

type, timing) of the transmitted signal to represent information.

Concentration-Based Techniques

The main idea of concentration-based techniques is carrying information on varying released MM concentration over fixed period discrete time slots (i.e., symbol durations, t_s). In its simplest form, each symbol represents one bit value and is called on-off keying (OOK). In OOK, the transmitter releases a fixed number of MMs (i.e., n_1) if the corresponding bit value (k -th symbol $S[k]$) is bit-1 or releases no molecules at all if it is bit-0 Mahfuz et al. (2010). At the receiver side, the receiver counts the number of MMs arrive within each symbol duration (i.e., $N^{Rx}[k]$) and makes a threshold-based decision to decode the bit value of the given symbol duration ($\hat{S}[k]$).

A more generalized version of OOK is called concentration shift keying (CSK) where each symbol, depending on the system design, represents m -bits of information (Kuran et al. 2011). From a communication point of view, CSK is akin to amplitude modulation (AM) method in analog modulation and amplitude shift keying (ASK) in digital modulation. In CSK, for the k th symbol in the message, the transmitter releases $N^{Tx}[k]$ number of MMs depending on the current symbol value as

$$N^{Tx}[k] = n_{S[k]}, \quad S[k] \in (0, 1, \dots, 2^m - 1), \quad (1)$$

where $S[k]$ can take one of the 2^m symbol values (e.g., $\text{sym}_0, \text{sym}_1$) in the modulation alphabet and $n_{S[k]}$ denotes the number of molecules to emit for the symbol $S[k]$. In order to decode $\hat{S}[k]$ from the received signal, the receiver uses $2^m - 1$ thresholds (i.e., $\lambda_0, \lambda_1, \dots, \lambda_{2^m-2}$) as

$$\hat{S}[k] = \begin{cases} \text{sym}_0, & N^{Rx}[k] < \lambda_0 \\ \text{sym}_i, & \lambda_i \leq N^{Rx}[k] < \lambda_{i+1} \\ \text{sym}_{2^m-1}, & \lambda_{2^m-2} \leq N^{Rx}[k] \end{cases} \quad (2)$$

Type-Based Techniques

A second group of modulation techniques, called type-based techniques, focuses on using multiple types of MMs in the communication system as a basis of a modulation technique. In these techniques, the transmitter has the capability of releasing different types of MMs (mm_{type}) that can only be received by a particular type of receptors at the receiver surface.

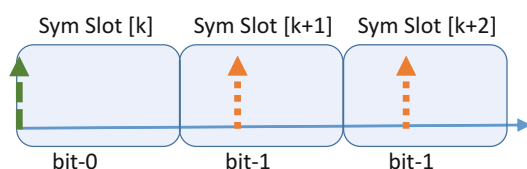
In the first type-based modulation technique, molecular shift keying (MoSK), each different symbol value (sym_i) is represented by a specific type of MM (Kuran et al. 2011; Aminian et al. 2015; Galmés and Atakan 2016). During each symbol duration, the receiver counts the number of arriving MMs for each MM type and decodes $\hat{S}[k]$ based on thresholding or the majority decision:

$$\hat{S}[k] = \text{sym}_i, \quad \text{where } i = \arg \max_r N_{mm_r}^{Rx}[k]. \quad (3)$$

Compared to CSK, MoSK considerably reduces the intersymbol interference (ISI) element of the signal due to the fact that each bit value is represented by a different type of molecule. On the other hand, it increases the system complexity since the transmitter is required to synthesize different types of molecules and the receiver is required to have multiple receptors on its surface. Moreover, molecule types should be similar in terms of diffusion coefficient due to propagation dynamics and single symbol duration.

Timing-Based Techniques

The main idea of timing-based modulations is to encode the information on the release time of the information molecules. In its simplest form, emission of MMs is done at an instance in the symbol slot where the time instance can be between the start and the end of the symbol slot. The time instance represents the intended symbol, and the receiver aims to detect the emission time hence the intended symbol. Capacity limit of timing-based techniques is derived and analyzed in Farsad et al. (2016a) and Murin et al. (2016).



Modulation in Molecular Signaling, Fig. 1 Example scenario showing the emission times with the bit sequence 011

Timing-based approach is also called pulse position modulation (PPM), which is similar to PPM in optical communications (Garraalda et al. 2011). In the binary case of timing-based techniques, the transmitter simply emits molecules at t_{sym_0} for sym_0 (which is usually at the start of the symbol duration) and emits molecules at t_{sym_1} for sym_1 (which is usually around the mid symbol duration) (Fig. 1).

Another release timing-based modulation scheme is called time-elapse communication (TEC) and is proposed for slow networks such as on-chip bacterial communication (Krishnaswamy et al. 2013). In TEC, information is encoded in the time interval between two consecutive pulses (i.e., two consecutive pulses of information particle transmission), which is similar to pulse-width modulation (PWM). At the receiver side, the information is decoded by measuring the duration between two pulses. The authors showed that TEC can outperform on-off keying when techniques such as differential coding are used.

Hybrid Techniques

The three modulation techniques addressed above constitute the main categories of modulation techniques for MC. These techniques, which utilize the molecular concentration, type, or release timing, are simple and easy to implement. They show low system complexity unless the modulation order goes up to very high levels. However, in practice more efficient ways to enhance performance depending on the system conditions are required. This subsection introduces more advanced techniques most of which are operated by combining two or more basic principles.

In a molecular concentration-based modulation system, one of the main issues to be solved is ISI. Since the same type of MM is used over multiple symbol duration, previously transmitted molecules affect the current symbol, which act as interference. Thus, there have been many studies aiming at effectively eliminating the ISI. The following techniques merge the advantages of CSK and MoSK: molecular concentration shift keying (MCSK) (Arjmandi et al. 2013), molecular transition shift keying (MTSK) (Tepekule et al. 2014), and zebra-CSK (Pudasaini et al. 2014).

Arjmandi et al. proposed a hybrid technique called MCSK, where if $S[k] = \text{sym}_1$, the transmitter releases (mm_a) type of MMs if it is an odd-numbered symbol or releases (mm_b) type of MMs if it is an even-numbered symbol Arjmandi et al. (2013). Similar to binary CSK, in case $S[k] = \text{sym}_0$, the transmitter does not release any MMs. The receiver decodes the signal by checking the concentration of (mm_a) or (mm_b) depending on the symbol parity. In the binary case, MCSK outperforms MoSK in terms of BER due to the fact that the ISI component of the molecular signal after a sequence of sym_1 's grows too large so that it is highly likely that the next symbol with sym_0 will be mis-decoded as sym_1 . The alternating approach of MCSK remedies such mis-decodings. However, the advantage of MCSK diminishes in the quadruple case consequently, quadruple MoSK outperforms quadruple MCSK.

Tepekule et al. proposed a hybrid technique called MTSK to eliminate the destructive ISI while utilizing the constructive ISI (Tepekule et al. 2014). The MTSK is also similar to MCSK, but it determines the molecule type to be transmitted considering the current and the previous symbols. MTSK changes the molecule type at the symbol transitions, which results in reduction of the destructive ISI.

A short comparison of all the modulation techniques mentioned in this entry is given in the following table based on their modulation technique group, receiver type, environment considered, and the performance metric used in the evaluation of the described technique

Modulation in Molecular Signaling, Table 1 Comparison of modulation techniques for molecular communication

Technique name	References	Group	Receiver type	Metrics used
CSK	Kuran et al. (2010, 2011), Mahfuz et al. (2011), and Lin et al. (2012)	Concentration	Abso-FPP	System Resp., Ch. capacity, SER
MoSK	Kuran et al. (2011), Aminian et al. (2015), and Galmés and Atakan (2016)	Type	Abso-FPP	Ch. capacity
Timing	Farsad et al. (2016a) and Murin et al. (2016)	Timing	Abso-FPP	Ch. capacity, BER
PPM	Garraalda et al. (2011)	Timing	Passive	System Resp.
TEC	Krishnaswamy et al. (2013)	Timing	Microfluidic	Ch. capacity
MCSK	Arjmandi et al. (2013)	Hybrid	Abso-FPP	BER
MTSK	Tepekule et al. (2014)	Hybrid	Abso-FPP	BER
Zebra-CSK	Pudasaini et al. (2014)	Hybrid	Abso-FPP	Ch. capacity

(Table 1). As for the receiver types, Abso-FPP refers to an absorbing receiver design which considers the first passage probability (FPP) of the MMs, whereas passive refers to a passive receiver in which upon coming into contact with the receiver, the MMs are not absorbed but are still allowed to move around in the environment.

Key Applications

The potential applications of MC include medical applications, control of chemical reactions, coordination among nanorobots in microscales, underground communications, underwater localization and communications, environmental monitoring, communication applications in pipes or duct systems, and coordination among search and rescue robots for harsh environments in macroscales.

Cross-References

- [Molecular Communication for Wireless Body Area Networks](#)
- [Nanonetworks](#)
- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

References

Abadal S, Akyildiz IF (2011) Automata modeling of quorum sensing for nanocommunication networks. *Elsevier Nano Commun Netw* 2(1):74–83

Aminian G, Mirmohseni M, Kenari MN, Fekri F (2015) On the capacity of level and type modulations in molecular communication with ligand receptors. In: *IEEE international symposium on information theory (ISIT)*, pp 1951–1955

Arjmandi H, Gohari A, Kenari MN, Bateni F (2013) Diffusion-based nanonetworking: A new modulation technique and performance analysis. *IEEE Commun Lett* 17(4):645–648

Cobo LC, Akyildiz IF (2010) Bacteria-based communication in nanonetworks. *Elsevier Nano Commun Netw* 1(4):244–256

Farsad N, Murin Y, Eckford A, Goldsmith A (2016a) On the capacity of diffusion-based molecular timing channels. In: *IEEE international symposium on information theory (ISIT)*, pp 1023–1027

Farsad N, Yilmaz HB, Eckford A, Chae CB, Guo W (2016b) A comprehensive survey of recent advancements in molecular communication. *IEEE Commun Surv Tut* 18(3):1887–1919

Feynman RP (1959) Plenty of room at the bottom. *Eng Sci (Caltech)* 23(5):22–36

Galmés S, Atakan B (2016) Performance analysis of diffusion-based molecular communications with memory. *IEEE Trans Commun* 64(9):3786–3793

Garraalda N, Llatser I, Cabellos-Aparicio A, Alarcón E, Pierobon M (2011) Diffusion-based physical channel identification in molecular nanonetworks. *Elsevier Nano Commun Netw* 2(4):196–204

Genc G, Kara YE, Yilmaz HB, Tugcu T (2016) ISI-Aware modeling and achievable rate analysis of the diffusion channel. *IEEE Commun Lett* 20(9):1729–1732

Krishnaswamy B, Austin CM, Bardill JP, Russakow D, Holst GL, Hammer BK, Forest CR, Sivakumar R (2013) Time-elapse communication: bacterial commu-

- nication on a microfluidic chip. *IEEE Trans Commun* 61(12):5139–5151
- Kuran MS, Yilmaz HB, Tugcu T, Ozerman B (2010) Energy model for communication via diffusion in nanonetworks. *Elsevier Nano Commun Netw* 1(2): 86–95
- Kuran MS, Yilmaz HB, Tugcu T, Akyildiz IF (2011) Modulation techniques for communication via diffusion in nanonetworks. In: *IEEE international conference on communication (ICC)*, Kyoto, pp 1–5
- Kuran MS, Tugcu T, Edis B (2012) Calcium signaling: overview and research directions of a molecular communication paradigm. *IEEE Wirel Commun* 19(5): 20–27
- Kuran MS, Yilmaz HB, Tugcu T (2013) A tunnel-based approach for signal shaping in molecular communication. In: *IEEE international conference on communication (ICC)*, Budapest, pp 776–781
- Lin WA, Lee YC, Yeh PC, Lee Ch (2012) Signal detection and ISI cancellation for quantity-based amplitude modulation in diffusion-based molecular communications. In: *2012 IEEE global communications conference (GLOBECOM)*, Anaheim, pp 4362–4367
- Mahfuz MU, Makrakis D, Mouftah HT (2010) On the characterization of binary concentration-encoded molecular communication in nanonetworks. *Elsevier Nano Commun Netw* 1(4):289–300
- Mahfuz MU, Makrakis D, Mouftah HT (2011) On the characteristics of concentration-encoded multi-level amplitude modulated unicast molecular communication. In: *2011 24th Canadian conference on electrical and computer engineering (CCECE)*, Niagara Falls, pp 312–316
- Murin Y, Farsad N, Chowdhury M, Goldsmith A (2016) On time-slotted communication over molecular timing channels. In: *Proceedings ACM international conference on nanoscale computer and communication*, New York, p 9
- Nakano T, Suda T, Moore M, Egashira R, Enomoto A, Arima K (2005) Molecular communication for nanomachines using intercellular calcium signaling. In: *Proceedings of IEEE international conference on nanotechnology (NANO)*, pp 478–481
- Nakano T, Eckford AW, Haraguchi T (2013) *Molecular communication*. Cambridge University Press, Cambridge
- Pudasaini S, Shin S, Kwak KS (2014) Robust modulation technique for diffusion-based molecular communication in nanonetworks. *arXiv preprint arXiv:1401.3938*
- Tepekule B, Pusane AE, Yilmaz HB, Tugcu T (2014) Energy efficient ISI mitigation for communication via diffusion. In: *IEEE international Black Sea conference on communication and networking (BlackSeaCom)*, Chişinău, pp 33–37
- Zimmer RK, Butman CA (2000) Chemical signaling processes in the marine environment. *Biol Bull* 198(2):168–187

Modulation Techniques for Synthetic Molecular Communications

► [Modulation in Molecular Signaling](#)

Molecular Abnormality Detection

► [Molecular Event Detection](#)

Molecular Anomaly Detection

► [Molecular Event Detection](#)

Molecular Bit Decisions

► [Molecular Bit Detection](#)

Molecular Bit Detection

Mark Leeson
University of Warwick, Coventry, UK

Synonyms

[Molecular bit decisions](#); [Molecular bit determination](#); [Molecular receiver decoding](#)

Definition

Molecular bit detection is the process at a molecular communications (MC) receiver of distinguishing between ones and zeros that have been sent by a transmitter.

Historical Background

Although Harry Nyquist is most often remembered for establishing 2 W pulses per second as the maximum signaling rate that can be supported over a baseband channel of bandwidth W, he also provided the axioms on which digital communications are built (Nyquist 1928). These are the division of time into defined intervals that are now referred to as symbol intervals and the conveying of information by changing a signal property in these intervals. The use of one bit of information per symbol leads to one binary digit (bit) being conveyed by each interval and to a bit rate that is purely the number of intervals per second. The binary digits must be put into a form that is matched to the transmission capabilities of the channel to be used. For most communication systems currently deployed, this entails modulating electromagnetic (EM) signals with the data to be transmitted leading to tractable analysis based on Maxwell's equations. Moreover, many established wired and wireless communication systems experience noise from thermal effects or other users that may be modeled using a Gaussian probability density function (PDF). This noise can cause problems in distinguishing a one bit from a zero bit causing bit errors. MC systems differ significantly from traditional communication systems in that they do not employ EM waves but rely on chemical carriers to convey their information. The natural world makes use of chemical signals for short- and long-range communications (Farsad et al. 2016). The biocompatibility and low energy nature of MC signals make them suitable for applications where EM signals cannot be employed. In general, MC may utilize molecular rails using carrier substances, may employ fluid flow, or may be based on diffusion (Deng et al. 2015). Here, the focus is on MC via Diffusion (MCvD) where molecules propagate randomly, making the process different to the spreading of EM waves. Despite the long evolutionary history of MCvD, it was only recently considered for microscale (Hiyama et al. 2005) and macroscale (Farsad et al. 2013) communication systems.

Foundations

The reliable recovery of data transmitted over a communication channel depends critically on the detection algorithm employed at the receiver. This estimates the signal from the noisy, distorted version that arrives after traversing a communication channel. In the case of bit detection, this will customarily entail the identification of a received bit as a "one" or a "zero." The most fundamental modulation scheme is referred to as on-off keying or OOK, where the presence of a signal in a bit period represents a "one" and signal absence a "zero." In the analysis that follows, this scheme is assumed for simplicity to illustrate the principle of bit detection. More sophisticated schemes have been developed for MCvD (Wang et al. 2017), and the analysis given below may be adapted accordingly. To achieve the goal of a communication system and transfer information from one location to another, a suitable signal must first be generated by the transmitter. This must then propagate to the receiver to be decoded. Thus, the system comprises transmitter, receiver, and channel. Here, the second two of these are particularly relevant since the receiver must make a decision on the received bits after they have suffered potential distortion and noise from the channel, which is the environment in which the signal propagates. In traditional communication systems, the channel is typically a wired or wireless connection carrying appropriate EM waves. In MCvD, particles act as chemical carriers to convey the information. These carriers are small, nanometers to micrometers in size, and travel in an aqueous or gaseous environment where they are free to move and diffuse. Noise may be introduced mainly by the channel through the diffusive nature of the carriers and the presence of other transmitters.

The received signal in MCvD is thus extremely random, and it is common to take the distribution of the received signal for a single molecule as the mean channel impulse response $\bar{h}(t)$. The number of molecules successfully received during the current bit interval or timeslot, N_0 , thus has mean given by:

$$\overline{N_0} = N_{tx} \overline{h}(t) \quad (1)$$

when the transmitter releases N_{tx} molecules to represent a one. There are several receiver models that have been proposed for different MCvD scenarios and each of these produces a different $\overline{h}(t)$. The principal categories are passive (Mahfuz et al. 2014), fully absorbing (Yilmaz et al. 2014), and reactive (Ahmadzadeh et al. 2016) depending of the behavior of the molecules and the receiver. The simplest form of $\overline{h}(t)$ results from a full absorbing receiver so that will be used as the basis for illustration here. The others produce results with similar behaviors but different optimum values and performance details.

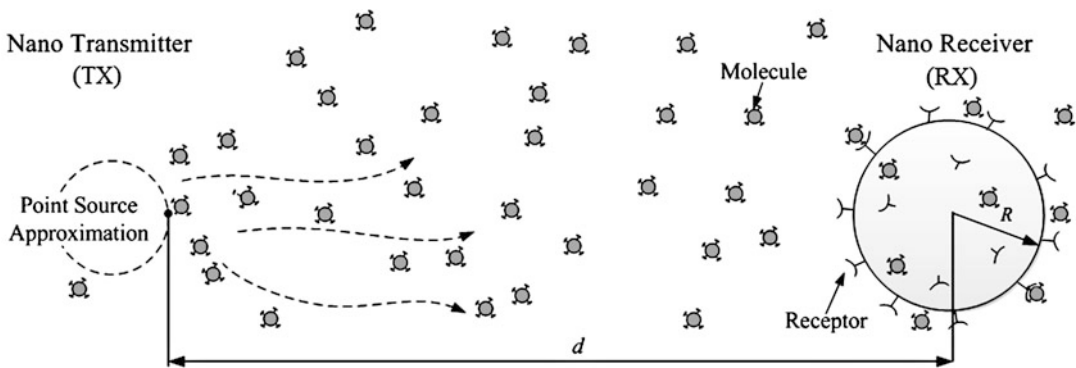
Although diffusion is a well-established topic in Physics, the solution of its defining equation in realistic scenarios for communicating information is very challenging. Therefore, it is common to make the approximation illustrated in Fig. 1 to the scenario where a point source transmitter of molecules (TX) releases N_{tx} of these at time zero and they travel to the receiver (RX) via 3D Brownian motion. A fraction of the molecules arrives at RX within the timeslot.

As introduced above, the simplest case (assumed here) is that RX is fully absorbing, and molecules are captured as soon as they collide with the surface of RX. The major impact of the different receiver types is to alter the probability $P_{ca}(d, t)$ that RX captures a molecule at a time t and distance d as defined in Fig. 1.

For full absorption and a counting receiver, the capture probability may be written as (Yilmaz et al. 2014):

$$P_{ca}(d, t) = \frac{R}{d} \operatorname{erfc} \left(\frac{d - R}{\sqrt{4Dt}} \right) \quad (2)$$

where R is the receiver radius and D is the diffusion coefficient in the medium in which communication is taking place. The fundamental representation of information is via a binary form with the transmitted information represented by a sequence of binary symbols. Each consecutive time slot contains either N_{tx} or no molecules. Each timeslot is of duration t_s and molecules arriving at RX are captured and removed from the environment. RX can either measure the number of molecules arriving at a certain time instant (amplitude detection) or the accumulated total of molecules arriving during a certain time period (energy detection) (Llatser et al. 2013). The latter is a better choice, particularly for long distance transmission, since it requires far fewer molecules to be transmitted for the same bit error rate (BER) (Aijaz and Aghvami 2015). Thus, it is assumed that RX determines whether a one or a zero was sent by counting the number of absorbed molecules. There have been developments in recent years to add further sophistication to the detection process within the constraints of the capabilities of nano-machines. These have included the detection of sequences of bits using both the maximum a posteriori (MAP) and maximum likelihood (ML) criteria (Kilinc and Akan



Molecular Bit Detection, Fig. 1 Molecular communication channel

2013; Meng et al. 2014). ML produces optimum receiver performance and offers viable suboptimum methods for nano-machines in a range of scenarios (Noel et al. 2014; Fang et al. 2018; Kuscü and Akan 2018).

To avoid detailed discussion of the ML method, for which readers are referred to the references above, single bit threshold detection is assumed. Thus, if the number of molecules received in an intended time slot exceeds a pre-designed threshold, the symbol is interpreted as “one,” otherwise, it is interpreted as “zero” (Kuran et al. 2011). It is possible to develop an adaptive threshold (Damrath and Hoehner 2016), but again for simplicity a fixed value is assumed here. The approximation is made that molecules are released as an impulse at the beginning of the timeslot and diffuse toward RX to be either registered or not registered by RX within the timeslot. Thus, the distribution is derived from a series of trials with the outcome received or not received producing a binomial distribution for N_0 :

$$N_0 \sim \text{Binom}(N_{\text{tx}}, P_{\text{ca},0}) \quad (3)$$

where $P_{\text{ca},0} = P_{\text{ca}}(d, t_s)$. For large numbers of molecules, this may be conveniently approximated by a normal distribution

$$N_{0_Norm} \sim \mathcal{N}(N_{\text{tx}} P_{\text{ca},0}, N_{\text{tx}} P_{\text{ca},0} \{1 - P_{\text{ca},0}\}) \quad (4)$$

but this can be inaccurate when the number of molecules is small (Lu et al. 2015, 2016), leading to an asymmetric binomial distribution. Thus, for N_{tx} values in the hundreds of molecules, the Poisson approximation may be utilized:

$$N_{0_Pois} \sim \text{Pois}(N_{\text{tx}} P_{\text{ca},0}). \quad (5)$$

However, the molecules released at TX are not guaranteed to reach RX in one timeslot, so the remaining molecules may arrive later causing intersymbol interference (ISI). The number of molecules received from the previous i^{th} symbol in the current time slot is represented by N_i , $i \in \{1, \dots, I\}$, where I is the ISI length (Lu et al.

2015).

$$N_i \sim \text{Binom}(N_{\text{tx}}, P_{\text{ca},i} - P_{\text{ca},i-1}) \quad (6)$$

$$P_{\text{ca},i} = P_{\text{ca}}(d, \{i+1\} t_s). \quad (7)$$

With corresponding approximations

$$N_{i_norm} \sim \mathcal{N}(N_{\text{tx}} \Delta P_i, N_{\text{tx}} \Delta P_i \{1 - \Delta P_i\}) \quad (8)$$

and

$$N_{i_Pois} \sim \text{Pois}(N_{\text{tx}} \Delta P_i) \quad (9)$$

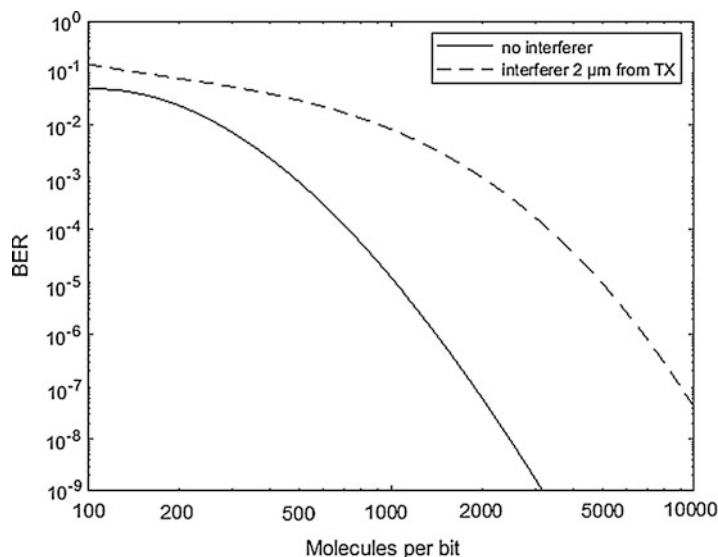
where $\Delta P_i = P_{\text{ca},i} - P_{\text{ca},i-1}$.

From these expressions, as shown in (Lu et al. 2017), it is possible to write down the effect of the number of previous slots that comprise the ISI length. Based on this analysis, a prediction of the MCvD system BER may be made by determining the probabilities that the noise sources will cause the signal at RX to be the wrong side of the decision level. Thus, in Fig. 2 the solid line indicates the BER obtained when the distance d is 7 μm and the diffusion coefficient D is 79.4 $\mu\text{m}^2\text{s}^{-1}$ (Lu et al. 2016).

In common with many other wireless systems, MCvD is also susceptible to one further source of error, namely interference from other users. This may be quantified by including an interfering receiver in the system that may absorb molecules destined for RX considered above (Lu et al. 2016). To illustrate a particularly problematic case, the dashed line in Fig. 2 illustrates the BER obtained over the original path when there is an interferer on the same side as the RX but at a distance of just 2 μm from TX (Lu et al. 2016). Although this is a particularly bad case that produces a power penalty of almost 7 dB at a BER level of 10^{-6} , it does graphically make the point that care must be taken when transmission is over more than just a point to point link. The use of different molecules for transmission to different receivers is possible (Kuran et al. 2011) but adds considerable complexity to MCvD systems.

Molecular Bit Detection,

Fig. 2 BER results for a MC system with 7 μm between TX and RX; solid line with no interferer; dotted line with an interferer 2 μm from TX



Improving Decisions

The reliability of information transmission can be increased by the employment of error correcting codes (ECCs) as is the case to improve the BER performance of conventional communication systems (Costello and Forney 2007). The small size and limited capabilities of nano-machines means that codes must be chosen for energy efficiency (Bai et al. 2014) and simplicity of implementation (Leeson and Higgins 2012). In recent years, considerable research efforts have been made in the development of nanoscale logic gates (Pilarczyk et al. 2018), making ECC implementation conceivable in the near future. ECC performance has thus been investigated numerically for MCvD in anticipation of future developments (Lu et al. 2015).

Key Applications

In areas where EM communication is not suitable, MC offers a solution for the near future. At the macroscale, this includes tunnel or pipe networks where experimental results have shown that EM waves fail to propagate but molecules get through but with a long delay (Qiu et al. 2014). This would facilitate, for example, the recovery of data

from embedded sensors or communications between search-and-rescue robots. Applications at the micro- or even nanoscale have driven the applications agenda to date with the promise of nanorobot drug delivery and tissue engineering (Nakano et al. 2013). In this regime, EM communication is difficult given the necessary ratio of the antenna size to wavelength ratio and the need for a line of sight using optical communications (Akyildiz et al. 2008).

Cross-References

- [Brownian Motion](#)
- [Modeling Approaches for Simulating Molecular Communications](#)
- [Modulation in Molecular Signaling](#)
- [Molecular Event Detection](#)
- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

References

- Ahmadzadeh A, Arjmandi H, Burkovski A, Schober R (2016) Comprehensive reactive receiver modeling for diffusive molecular communication systems: reversible binding, molecule degradation, and finite number of receptors. *IEEE Trans Nanobioscience* 15(7):713–727

- Aijaz A, Aghvami A-H (2015) Error performance of diffusion-based molecular communication using pulse-based modulation. *IEEE Trans Nanobioscience* 14(1):146–151
- Akyildiz IF, Brunetti F, Blazquez C (2008) Nanonetworks: a new communication paradigm. *Comput Netw* 52(12):2260–2279
- Bai C, Leeson MS, Higgins MD (2014) Minimum energy channel codes for molecular communications. *Electron Lett* 50(23):1669–1671
- Costello DJ, Forney GD (2007) Channel coding: the road to channel capacity. *Proc IEEE* 95(6):1150–1177
- Damrath M, Hoehner PA (2016) Low-complexity adaptive threshold detection for molecular communication. *IEEE Trans Nanobioscience* 15(3):200–208
- Deng Y, Noel A, Elkashlan M, Nallanathan A, Cheung KC (2015) Modeling and simulation of molecular communication systems with a reversible adsorption receiver. *IEEE Trans Mol Biol Multi-Scale Commun* 1(4):347–362
- Fang Y, Noel A, Yang N, Eckford AW, Kennedy RA (2018) Maximum likelihood detection for cooperative molecular communication. In: *Proceedings of the IEEE international conference on communications (ICC)*, Piscataway, NJ pp 1–7
- Farsad N, Guo W, Eckford AW (2013) Tabletop molecular communication: text messages through chemical signals. *PLoS One* 8(12):e82935
- Farsad N, Yilmaz HB, Eckford A, Chae C-B, Guo W (2016) A comprehensive survey of recent advancements in molecular communication. *IEEE Commun Surv Tutorials* 18(3):1887–1919
- Hiyama S, Moritani Y, Suda T, Egashira R, Enomoto A, Moore M, Nakano T (2005) Molecular communication. In: *Proceedings of the NSTI nanotechnology conference and trade show*, Austin, TX, 3, pp 391–394
- Kilinc D, Akan OB (2013) Receiver design for molecular communication. *IEEE J Sel Areas Commun* 31(12):705–714
- Kuran MS, Yilmaz HB, Tugcu, T, Akyildiz AF (2011) Modulation techniques for communication via diffusion in nanonetworks. In: *Proceedings of the IEEE international conference on communications (ICC)*, Piscataway, NJ pp 1–5
- Kuscu M, Akan OB (2018) Maximum likelihood detection with ligand receptors for diffusion-based molecular communications in internet of bio-nano things. *IEEE Trans Nanobioscience* 17(1):44–54
- Leeson MS, Higgins MD (2012) Forward error correction for molecular communications. *Nano Commun Netw* 3(3):161–167
- Llatser I, Cabellos-Aparicio A, Pierobon M, Alarcon E (2013) Detection techniques for diffusion-based molecular communication. *IEEE J Sel Areas Commun* 31(12):726–734
- Lu Y, Higgins MD, Leeson MS (2015) Comparison of channel coding schemes for molecular communications systems. *IEEE Trans Commun* 63(11):3991–4001
- Lu Y, Higgins MD, Noel A, Leeson MS, Chen Y (2016) The effect of two receivers on molecular communications systems. *IEEE Trans Nanobioscience* 15(8):891–900
- Lu Y, Higgins MD, Leeson MS, Chen Y, Jennings P (2017) A revised look at the effects of the channel model in molecular communication systems. *IET Micro Nano Lett* 12(2):136–139
- Mahfuz M, Makrakis D, Mouftah H (2014) A comprehensive study of sampling-based optimum signal detection in concentration-encoded molecular communication. *IEEE Trans Nanobioscience* 13(3):208–222
- Meng LS, Yeh P-C, Chen K-C, Akyildiz IF (2014) On receiver design for diffusion-based molecular communication. *IEEE Trans Signal Process* 62(22):6032–6044
- Nakano T, Eckford AW, Haraguchi T (2013) *Molecular communications*. Cambridge University Press. New York, NY
- Noel A, Cheung KC, Schober R (2014) Optimal receiver design for diffusive molecular communication with flow and additive noise. *IEEE Trans Nanobioscience* 13(3):350–362
- Nyquist H (1928) Certain topics in telegraph transmission theory. *Trans AIEE* 47(2):617–644
- Pilarczyk K, Wlaźlaka E, Przyczyna D, Blachecki A, Podborska A, Anathasiou V, Konkoli Z, Szaciłowski K (2018) Molecules, semiconductors, light and information: towards future sensing and computing paradigms. *Coord Chem Rev* 365:23–40
- Qiu S, Guo W, Wang S, Farsad N, Eckford A (2014) A molecular communication link for monitoring in confined environments. In: *Proceedings of the IEEE international conference on communications (ICC)*, Piscataway, NJ pp 718–723
- Wang J, Yin B, Peng M (2017) Diffusion based molecular communication: principle, key technologies, and challenges. *China Commun* 14(2):1–18
- Yilmaz HB, Heren AC, Tugcu T, Chae C-B (2014) Three-dimensional channel characteristics for molecular communications with an absorbing receiver. *IEEE Commun Lett* 18(6):929–932

Molecular Bit Determination

► Molecular Bit Detection

Molecular Communication for Wireless Body Area Networks

M. D. Nashid Anjum and Honggang Wang
Department of Electrical Engineering,
University of Massachusetts Dartmouth, North
Dartmouth, MA, USA

Acronyms

BAN	Body Area Network
BANN	Body Area Nanonetwork
BAN ²	Body Area Nanonetwork
MC	Molecular Communication
IoNT	Internet of Nano Things
WBAN	Wireless Body Area Network

Definition

A wireless body area network (i.e., WBAN) generally consists of a central hub (i.e., base station) and a number of lightweight, low-powered, miniature, wearable sensors that operate in the proximity of a human body. Typically, it is located on the body, garment, underneath the skin, or deep into the tissue (Cavallari et al. 2014; Anjum and Wang 2016). WBAN is also known as a wireless body sensor network (WSN) or simply a body area network (BAN). On the contrary, a molecular communication (i.e., MC) system is an in-body BAN which consists of bio-nanomechanics such as biosensors and bio-actuators where the data transmission between the transmitter and receiver nanomachines is carried on by the encoded molecules (Cavallari et al. 2014; Nakano et al. 2012; Akyildiz et al. 2008). MC for WBAN is also termed as body area nanonetworks, i.e., BANN or BAN² (Suzuki et al. 2014; Atakan et al. 2012). The integration of MC-based in-body BANN and wearable WBAN is known as the Internet of Nano Things (IoNT) Dressler et al. (2015).

Historical Background

The use of the term “molecular communication” can be traced back to the 1970s where a number of researchers individually studied the various modes of molecular transmission among neuron cells (Dismukes and Key 1979). Later on, MC in host-parasite interaction and MC between plant-pathogen interaction are studied in Dixon and Lamb (1990) and Hadwiger et al. (1981), respectively. However, in the field of communication engineering, the concept of MC was developed in the early 2000s to create a communication network for bio-nanomachines (Nakano et al. 2005, 2012).

System Architecture

The system architecture of a MC typically consists of four key components – the sender, receiver, information bearing molecules (also known as signaling molecule), and propagation medium. To support the movement of the information molecule (i.e., IM), a MC may include an interface, transport, guide, and addressing molecules. All of these components are various forms of bio-nanomachines which are biochemically reactive, micrometer to nanometer range devices. The transfer of information in a typical MC takes the following five steps: encoding, transmitting, propagation, receiving, and decoding.

- *Encoding and Transmitting:* The encoding and transmitting of IM are done by the transmitter bio-nanomachine which is also known as the sender. The sender encodes the information in various forms within the molecules, such as the type of molecule used, their 3D structure (e.g., protein structure), sequential structure (e.g., DNA sequence), concentration (e.g., calcium concentration), number of molecule, release time of molecules, etc. The release of IM into the propagation medium is done

by budding vesicles (if the sender is a cell), opening the molecular gate (e.g., ion channel) of the sender membrane, or catalyzing a chemical reaction (Nakano et al. 2013).

- *Propagation:* The scales of propagation of the CM are divided in three categories – intracellular ($\leq 100 \mu\text{m}$), intercellular ($\leq 100 \mu\text{m}$), and inter-organ (up to few meters). The mode of propagation can be two types – passive and active. In the passive mode, a large number of molecules diffuse in all available directions. This is a slower process and does not require any infrastructure hence suitable for intracellular and intercellular communication. In addition, the propagation time t over a distance L is $t \approx \frac{L^2}{D}$, where D is the diffusion coefficient which depends on the molecular size and structure, viscosity of the medium, and the temperature. If $D = 100 \mu\text{m}^2/\text{s}$, a molecule takes about 2.78 h to propagate over a distance of 1 mm. Unlike a passive mode, the active mode of propagation directs the movements of molecules toward the targeted location. It is comparatively a faster process and takes fewer molecules thus suitable for inter-organ propagation scale. However, unlike passive modes, it requires propagation infrastructure and a constant supply of energy. Different forms of active transportation provide different propagation speeds such as motor proteins (e.g., kinesin, $2\text{--}4 \mu\text{m}/\text{s}$), bacterial chemotaxis (e.g., *E. coli*, few $\mu\text{m}/\text{s}$), and diffusion-reaction mechanism (e.g., calcium signaling, $20 \mu\text{m}/\text{s}$), etc. (Nakano et al. 2013). However, to support the information molecules, the propagation medium may contain some other types of molecules, such as transport molecules, to move information molecules, guide molecules to direct the movement of the transport molecules, interface molecules to selectively transport the information molecules, and address molecules to specify the receiver (Nakano et al. 2013).
- *Receiving and Decoding:* Receiving and decoding processes are done by the receiver nanomachine. Reception of information

molecules can be done through permeable plasma membrane, surface receptors, or chemically gated surface channels. The decoding process is essentially the reaction with the information molecule upon the reception at the receiving end. Furthermore, the consequences of the chemical reaction during decoding may generate movement, morphological changes, change in chemical functionality, or new molecules. The receiver may initiate multihop communication by releasing the produced molecules into the propagation medium (Nakano et al. 2013).

Communication Engineering Aspects

Although MC exists in nature for billions of years, the advancement in MC engineering has started primarily from the early 2000s. However, compared to traditional wireless communication, the advancement of MC is still in infancy (Farsad et al. 2016). This section sheds light on some of the engineering aspects of MC.

Physical Layer Aspects

The physical layer deals with the biophysical mechanism of data transmission, propagation, and reception. In essence, the data is transmitted in the form of IMs, also known as signal molecules. Moreover, the physical layer also provides the biophysical basis of hardware and interface (Farsad et al. 2016; Nakano et al. 2013).

- *Hardware:* The key hardware for MC are the sender, propagation channel, receiver, IM, and supporting molecules, as mentioned in Section “[Applications](#).” Essentially, these are all various kinds of bio-nanomachines.
- *Data/Message:* In a traditional WBAN, the physical layer data is transmitted into the form of 2.4 or 60 GHz mmWave (Anjum et al. 2017). Unlike traditional WBAN, the physical layer data or messages of MC are transmitted in the form of information molecules which

can be various kinds of protein molecules, vesicles, or bacteria.

- **Modulation Techniques:** Modulation in MC is the process of encoding data/messages in the IMs. The modulation or encoding of IMs can be done in various ways as we discussed in section “Applications.”
- **Propagation Channel Models:** Propagation channel models for intercellular and intracellular MC are known as microscale models. On the other hand, propagation channel models for inter-organisms are known as macroscale models. Major passive propagation models for microscale MC are free diffusion propagation and diffusion with first fitting. Major active propagation models are flow-assisted propagation, bacteria-assisted propagation, motor protein moving over microtubule tracks, microtubule filament motility over stationary kinesin, neurochemical propagation, and propagation through gap junction. The major macroscale propagation models are diffusion and flow-based propagation. Some other macroscale propagation models are advection, convection, mechanical dispersion, and turbulent flows. Macroscale models require comparatively larger amount of molecules and external source of power; hence, these models are active in nature (Farsad et al. 2016).
- **Noise Sources:** The sources of noise are primarily transmitter emission noise, random propagation, i.e., diffusion noise, reception noise, environmental noise such as degradation and reaction, multiple transmitters, etc.
- **Channel Capacity:** Practically, MC channels are not memoryless; thus, the channel capacity is different for different models of propagation. If x^n is a sequence of n successive transmission symbols, and y^n is the corresponding received symbols, then the channel capacity can be presented by Farsad et al. (2016):

$$\mathcal{C} = \liminf_{n \rightarrow \infty} \sup_{x^n} I(X^n; Y^n) \quad (1)$$

Link Layer Aspects

- **Error Handling:** To detect the transmitted message, the number of received molecules must be larger than a threshold. Hence, transmitting larger numbers of molecules increases the signal to interference plus noise ratio. Another approach of avoiding error is optimizing the transmission rate. Error detection and correction can be done by adding redundant information to the information molecule. For instance, adding an additional DNA sequence to a DNA molecule allows error detection and forward error correction at the receiving end (Nakano et al. 2012).
- **Addressing:** Addressing is the process of specifying the target receiver. Usually, addressing of the receiver is done by the type of molecule transmitted by the sender.
- **Synchronization:** Synchronization of the sender and receiver in MC is a complicated process because of the immense propagation delay and jitter. No significant research regarding synchronization is reported so far.
- **Media Access Control:** The interference among multiple pairs of senders and receivers can be avoided by transmitting different types of molecules for each sender-receiver pair. This technique is not sufficient for a large number of sender-receiver pairs. In that case, some other techniques could be employed such as channel reservation, switching or time-division multiplexing, etc. However, such approaches are not implemented yet.
- **Flow Control:** If the chemical reaction on the receiver side is significantly slower compared to the sender side, then the flow control is required to avoid buffer overflow on the receiver side. No such work has been reported yet in the literature.

Network Layer Aspects

The routing of the existing MC is limited to the static routing tables which fails to address dynamic locations of the network nodes. In case of a static routing, a sender transmits the data using a bacterium with addressing

molecules which indicates the target receiver. The router node receives the bacterium and employs statically defined chemical processes to retransmit the data using a bacterium which follows the chemical gradients to the following target router. Advancements regarding other network layer issues such as congestion control, topology management, network scalability, etc. are still limited (Nakano et al. 2013).

Mathematical Models

The simplest mathematical model to represent the movement of the information molecule is a random walk model which does not consider the directional drift or chemical reaction. Hence, a random walk model is suitable for representing passive propagation models. A more complex mathematical model is a random walk with drift which is suitable for representing the active propagation models where the movement of the IM is drifted directionally. Another kind of mathematical model for MC is the random walk with chemical reactions. This model considers the chemical reactions of IMs induced by the amplifiers. Amplifiers are located in the environment and react with molecules that can increase the reliability of the molecular propagation by multiplying the number of propagating molecules.

Applications

Although the primary application of molecular communication evolves around biomedical fields, its application in environmental and manufacturing fields is also reported (Nakano et al. 2012).

- *Biomedical Applications:* In the biomedical field, the applications of MC are health monitoring within an organism, disease diagnosis known as lab-on-a-chip, drug delivery within an organism, regenerative medicine, etc.
- *Environmental Applications:* MC can be used for monitoring and controlling environmental pollution.
- *Manufacturing Applications:* MC can be exploited to control the transport of bio-

nanomachines or molecules and can also be modified to develop novel patterns and structures of molecules (Nakano et al. 2013).

Conclusions

Although the concept of BAN² has been introduced in the early 2000s, the advancement of this research field is still at its infancy level, after two decades. However, BAN² has an immense potential and prospect in the biomedical and environmental applications. As a result, IEEE P1906.1 Standards Working Group for Nanonetworking has been formed in 2011 for standardizing the MC protocols.

References

- Akyildiz IF, Brunetti F, Blázquez C (2008) Nanonetworks: a new communication paradigm. *Comput Netw* 52(12):2260–2279
- Anjum MN, Wang H (2016) Optimal resource allocation for deeply overlapped self-coexisting WBANs. In: 2016 IEEE global communications conference (GLOBECOM). IEEE
- Anjum MDN, Fang H (2017) Coexistence in millimeter-wave WBAN: a game theoretic approach. In: 2017 international conference on computing, networking and communications (ICNC). IEEE
- Atakan B, Akan OB, Balasubramaniam S (2012) Body area nanonetworks with molecular communications in nanomedicine. *IEEE Commun Mag* 50(1):28–34
- Cavallari R et al (2014) A survey on wireless body area networks: technologies and design challenges. *IEEE Commun Surv Tutor* 16(3):1635–1657
- Dismukes RK (1979) New concepts of molecular communication among neurons. *Behav Brain Sci* 2(3):409–416
- Dixon RA, Lamb CJ (1990) Molecular communication in interactions between plants and microbial pathogens. *Annu Rev Plant Biol* 41(1):339–367
- Dressler F, Fischer S (2015) Connecting in-body nano communication with body area networks: challenges and opportunities of the internet of nano things. *Nano Commun Netw* 6(2):29–38
- Farsad N et al (2016) A comprehensive survey of recent advancements in molecular communication. *IEEE Commun Surv Tutor* 18(3):1887–1919
- Hadwiger LA, Loschke DC (1981) Molecular communication in host-parasite interactions: hexosamine polymers(Chitosan) as regulator compounds in race-specific and other interactions. *Phytopathology* 71(7):756–762

- Nakano T et al (2005) Molecular communication for nanomachines using intercellular calcium signaling. In: 5th IEEE conference on nanotechnology. IEEE
- Nakano T et al (2012) Molecular communication and networking: opportunities and challenges. *IEEE Trans Nanobioscience* 11(2):135–148
- Nakano T, Eckford AW, Haraguchi T (2013) *Molecular communication*. Cambridge University Press, Cambridge
- Suzuki J et al (2014) A service-oriented architecture for body area nanonetworks with neuron-based molecular communication. *Mob Netw Appl* 19(6):707–717

Molecular Communication Simulators

- [Modeling Approaches for Simulating Molecular Communications](#)

Molecular Event Detection

Reza Mosayebi
Sharif University of Technology, Tehran, Iran

Synonyms

[Molecular abnormality detection](#); [Molecular anomaly detection](#); [Molecular target detection](#)

Definitions

Molecular event detection is one of the key challenges in many microscale revolutionary applications, such as environmental and health monitoring and disease diagnosis, which deals with detecting undesired signals at the nano- and microscales.

Historical Background

One of the key challenges in disease diagnosis and health monitoring applications is the problem of event detection, e.g., early tumor detection, which has received significant attention in

medicine and other related fields (Chen et al. 2016; Wuab and Qu 2015). Upon detecting tumor (event), a drug can be released at an appropriate rate to mitigate the effect of tumor (Chude-Okonkwo et al. 2017). This motivates the investigation of event detection at microscale using molecular communications.

Event detection has been extensively studied in different fields (Chandola et al. 2009). In this context, event is also referred to as abnormality, anomaly, and outlier which has several applications including failure detection in computer science, segmentation of signals in biomedical applications, and fraud detection for credit cards. However, in molecular communication systems, due to diffusion of molecules, the number of molecules received at a receiver (received signal) has Poisson distribution which imposes inherent randomness to the molecular channels. Due to this specific characteristic of molecular communication, classical methods cannot be directly applied for detecting molecular events. Therefore, several approaches have been proposed for molecular event detection (Felicetti et al. 2014; Okaie et al. 2016; Ghavami and Lahouti 2017; Nakano et al. 2017; Mai et al. 2017; Mosayebi et al. 2017). In all of the proposed works for molecular event detection, multiple nanosensors (NSs) for initial detection of event are employed. Upon detection, the NSs send their decisions via releasing an amount of molecules to a fusion center (FC) where the final decision regarding the presence of event is made. In particular, in Felicetti et al. (2014), it is assumed that mobile NSs move through the vasculature and gather at the target location by binding to a tumor. Next, in Okaie et al. (2016), a similar idea is used where two types of NSs are used, named leader and follower NSs. Here, leader NSs create an attractant gradient around the target. Then, follower NSs move according to the attractant gradient, approach the target, and perform necessary actions such as releasing drug. Employing static NSs is proposed for anomaly detection in body tissue in Ghavami and Lahouti (2017), Mai et al. (2017), and Mosayebi et al. (2017). In particular, in Ghavami and Lahouti (2017) a macroscale noise channel is considered between

the NSs and the FC, while in Mai et al. (2017) and Mosayebi et al. (2017), a microscale Poisson signal-dependent noise channel is considered. Finally, an abstract modeling and simulation of both static and mobile NS network for target detection is proposed in Nakano et al. (2017).

System Model

For molecular event detection, one can consider a network of M NSs that observe a part of a tissue or move inside the vasculature and sense their environment (sensing scheme), along with an FC. Upon detecting an event (target or abnormality), NSs will release an amount of molecules into the medium, where some of them may be collected by the FC (reporting scheme) and where the final decision regarding the presence of an event will be made. The absence and presence of event are denoted by hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. In the following, the sensing and reporting schemes will be briefly described. Before going into the details of sensing and reporting schemes, it is noteworthy to mention that the approaches that will be described in the following are examples of what has been done in the molecular communication literature and are non-exhaustive for what might be practical.

Sensing Model

Each NS can measure one or more sensing variables, referred to as inputs. For instance, for tumor detection, examples of inputs include the concentration of specific types of molecules, referred to as *biomarkers*, such as nucleic acids and proteins, and a lack of oxygen (Wuab and Qu 2015; Chude-Okonkwo et al. 2017). Then, based on the measured inputs, the NS determines a numerical value which reflects the presence or absence of the event. For the sensing scheme, there are two different models in the literature, referred to as *hard* and *soft* decision schemes. For the hard decision scheme (Nakano et al. 2017; Ghavami and Lahouti 2017; Mai et al. 2017), the decision made by each NS $k \in \{1, \dots, M\}$ is a Bernoulli random variable (RV) C_k with the realization values of $c_k = 0$ and $c_k = 1$ upon

detecting hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. However, for the soft decision scheme, one of the $L > 2$ values in $\mathcal{X} = \{0, 1/(L-1), 2/(L-1), \dots, 1\}$ can be assumed as a soft decision for the k -th NS, where small and large values of c_k indicate that NS k leans toward hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively (Mosayebi et al. 2017). For RV C_k under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, general probability mass functions (PMFs) $g_0(\cdot)$ and $g_1(\cdot)$, respectively, can be assumed, where under hard decision, i.e., $L = 2$, they have only two nonzero values for $c_k = 0$ and $c_k = 1$. It is noteworthy to indicate that the functionality of $g_i(\cdot)$, $i = 0, 1$ with respect to its argument is a function of several factors such as the NS structure and its sensing accuracy.

Reporting Model

For each sensor k , if $c_k \neq 0$, an instantaneously number (Ghavami and Lahouti 2017; Mai et al. 2017; Mosayebi et al. 2017) or a constant rate of molecules which can be degraded in the medium (Okaie et al. 2016; Nakano et al. 2017) will be released toward the FC. For each of the released schemes, a mean value of $g_i(c_k)\theta_k$ for the number of molecules released by the k -th NS and collected at the FC can be considered. Here, θ_k is a function of the number/rate of the released molecules, distance between the NS and the FC, the diffusion coefficient of the released molecules, and the shape and size of the FC. In line with the references for molecular event detection and for tractability of analysis, in the following it is assumed that $\theta_k = \theta, \forall k \in \{1, \dots, M\}$. Here, two cases can be imagined:

1. (Case I) Each NS sends a unique type of molecules, different from the molecules released by other NSs.
2. (Case II) The same type of molecule is used by all NSs.

In the following, for each case, the PMF of the signal received at the FC is modeled. Derived PMFs for the signal received at the FC will be further used to design optimal decision rule for the FC.

Case I

For Case I, since the types of molecules released by different NSs are different, the FC can distinguish between the messages sent by different NSs. Furthermore, an environmental noise for each type of molecules can be considered which is an independent source of molecules. Let Y_k denote the RV which models the number of released molecules by the k -th NS and the corresponding environmental noise present at the FC. Therefore, conditioned on hypothesis $\mathcal{H}_i, i = 0, 1$, the PMF of Y_k can be modeled as follows (Mosayebi et al. 2017, Eq. (8)):

$$P(Y_k = y_k | \mathcal{H}_i) = \sum_{c_k \in \mathcal{X}} g_i(c_k) \times \quad (1)$$

$$\frac{\exp(-g_i(c_k)\theta - \lambda_k) (g_i(c_k)\theta + \lambda_k)^{y_k}}{y_k!}, \quad (2)$$

$i = 0, 1,$

where $P(\cdot)$ denotes probability measure, y_k is a realization of Y_k , and λ_i is the mean number of the environmental noise molecules with the same type as the molecules released by NS k . Since the types of molecules are different, the RVs Y_k become independent. Now, by defining $\mathbf{Y} = [Y_1, \dots, Y_M]^T$ as the vector containing all RVs $Y_k, k \in \{1, \dots, M\}$, the PMF of the signal received at the FC can be characterized as follows:

$$P(\mathbf{Y} = \mathbf{y} | \mathcal{H}_i) = \prod_{k=1}^M P(Y_k = y_k | \mathcal{H}_i) \quad (3)$$

$$= \prod_{k=1}^M \sum_{c_k \in \mathcal{X}} g_i(c_k) \times \frac{\exp(-g_i(c_k)\theta - \lambda_k) (g_i(c_k)\theta + \lambda_k)^{y_k}}{y_k!},$$

$i = 0, 1,$

where $\mathbf{y} = [y_1, \dots, y_M]^T$ is a realization of $\mathbf{Y}$.

Case II

For Case II, since the types of molecules released by different NSs are the same, the FC cannot dis-

tinguish between the messages of different NSs. Therefore, the PMF of the number of molecules received at the FC is a function of RV $C = \sum_{k=1}^M C_k \in \mathcal{X} = \{0, 1/L - 1, 2/L - 1, M\}$. Since RVs C_k are statistically independent, the PMF of RV C can be obtained using convolution operator. This PMF is denoted by $G_0(\cdot)$ and $G_1(\cdot)$ for hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Furthermore, the same type of molecule can always be considered in the environment which plays the role of environmental noise. Therefore, by denoting the random variable Y as the total number of released molecules by different NSs and environmental noise that are collected at the FC, one can see that conditioned on hypothesis $\mathcal{H}_i, i = 0, 1$, Y can be modeled as follows (Mosayebi et al. 2017, Eq. (17)):

$$P(Y = y | \mathcal{H}_i) = \sum_{c \in \mathcal{X}} G_i(c) \times \quad (4)$$

$$\frac{\exp(-G_i(c)\theta - \lambda) (G_i(c)\theta + \lambda)^y}{y!}, \quad (5)$$

$i = 0, 1,$

where y is the realization of Y , c is the realization of C , and λ is mean number of environmental noise molecules.

Detector Design for the FC

For the FC, a decision $d \in \{0, 1\}$ is made, where $d = 0$ and $d = 1$ means that the FC chooses hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Due to the inherent randomness of the received signal and the presence of noise, the selected hypothesis by the FC may be incorrect. Therefore, two types of errors can be imagined: first, if the FC chooses $d = 1$, while hypothesis $\mathcal{H}_0$ is correct. For this case, the conditional probability of error is referred to as *probability of false alarm* and is denoted by $P_{fa} = P(d = 1 | \mathcal{H}_0)$. Similarly, another type of error exists where the FC chooses $d = 0$, while hypothesis $\mathcal{H}_1$ is correct. For this case, the conditional probability of error is referred to as *probability of missed detection* and is denoted by $P_m = P(d = 0 | \mathcal{H}_1)$.

For the FC, the goal is to design the optimal decision rule that minimizes the probability of missed detection subject to a pre-assigned upper bound ω on the probability of false alarm. That is,

$$\min_{\text{detectors}} P_m, \text{ subject to } P_{fa} \leq \omega. \quad (6)$$

The optimal solution of (6) is the Neyman-Pearson detector (Kay 1998), which compares the log-likelihood ratio (LLR) with the maximum threshold τ that ensures $P_{fa} \leq \omega$. The LLR for each mentioned cases above can be written as

$$\text{LLR} = \begin{cases} \sum_{k=1}^M \log \left(\frac{P(Y_k=y_k|\mathcal{H}_1)}{P(Y_k=y_k|\mathcal{H}_0)} \right); & \text{Case I,} \\ \log \left(\frac{P(Y=y|\mathcal{H}_1)}{P(Y=y|\mathcal{H}_0)} \right); & \text{Case II,} \end{cases} \quad (7)$$

where $P(Y_k = y_k|\mathcal{H}_i)$ and $P(Y = y|\mathcal{H}_i)$, $i = 0, 1$, are given in (2) and (5), respectively. Hence, the optimal decision rule can be written as

$$d_{\text{Optimal}} = \begin{cases} 0, & \text{if } \text{LLR} \leq \tau, \\ 1, & \text{if } \text{LLR} > \tau, \end{cases} \quad (8)$$

Remark 1 Since the general form of LLR in (7) may be computationally complex to be implemented in microscale, several approximations for LLR are proposed and investigated in Mosayebi et al. (2017).

Remark 2 In practice, for event detection, some system parameters such as θ and $g_i(\cdot)$, $i = 0, 1$ are not known for the FC. As a first step to address this, in Ghavami and Lahouti (2017), one unknown system parameter is assumed. For this case, in Ghavami and Lahouti (2017), the maximum likelihood estimation of the unknown system parameter is derived for the initial decision at the NSs. Then, based on the estimation, a suboptimal hard decision rule for the NSs is proposed and investigated.

Numerical Result

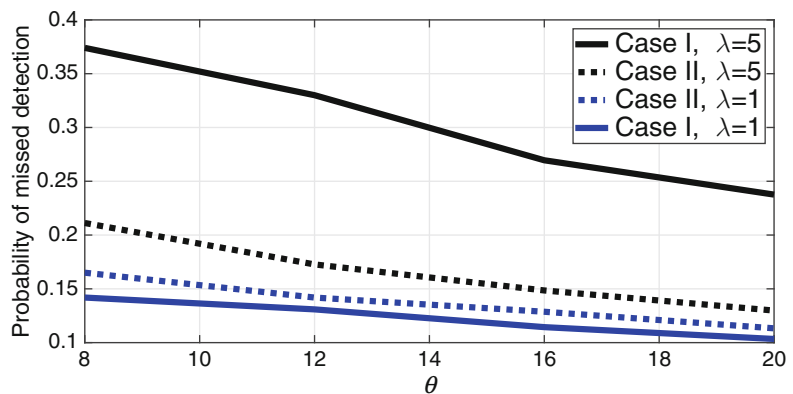
For a simple comparison between the proposed schemes, i.e., Cases I and II, the performance of Neyman-Pearson detectors is compared in Fig. 1 for

$$\begin{aligned} g_0(c_k) &= \frac{\exp(-2.5c_k)}{\sum_{x \in \mathcal{X}} \exp(-2.5x)}, \\ g_1(c_k) &= \frac{\exp(3.5c_k)}{\sum_{x \in \mathcal{X}} \exp(3.5x)}, \end{aligned} \quad (9)$$

$L = 4$, $M = 2$, and $P_{fa} = 0.05$. To this end, P_m is plotted versus θ for different values of λ . To have a fair comparison between Cases I and II, it is assumed that $\lambda_1 = \lambda_2 = \lambda$. From Fig. 1, it can be observed that as the value of θ increases, P_m decreases since the reporting scheme becomes more reliable. In addition, one can see that the relative performance of Case I and Case II depends on the value λ . For large λ ,

Molecular Event Detection, Fig. 1

Probability of missed detection versus θ for Cases I and II using the respective optimal decision rules for $P_{fa} = 0.05$. All curves were obtained via simulation



Case II outperforms Case I, whereas for small λ , Case I outperforms Case II. In fact, the advantage of Case II over Case I comes from the fact that the mean number of noise molecules for each type of molecule is assumed to be constant, i.e., $\lambda_1 = \lambda_2 = \lambda$, which leads to a higher overall noise for Case I compared to Case II. Alternatively, the trivial advantage of Case I over Case II is that the FC can distinguish between the molecules released by different NSs and exploit this additional knowledge for the improvement of the detection performance. Therefore, the superiority of Case I over Case II depends on the system parameters.

Future Directions

Although there are several works for molecular event detection, there exist some issues to be addressed. For instance, having a more adequate sensing model for particular applications such as event detection in blood vessels, body tissues, and air is highly recommended. The sensing model should be a function of the NS structure and the stochasticity of the event. In addition, considering more practical assumptions such as unknown sensing distribution, unknown locations for NSs, and unknown target location will result in different decision rules for the NSs and the FC. The idea of jointly estimation and detection which is introduced in the conventional communication systems can be also specialized for the molecular event detection systems.

Cross-References

- [Applications of Molecular Communication Systems](#)
- [Drug Delivery via Nanomachines](#)
- [Molecular Bit Detection](#)

References

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. *ACM Comput Surv* 41(3):1–58

- Chen G, Roy C, Prasad PN (2016) Nanochemistry and nanomedicine for nanoparticle-based diagnostics and therapy. *Chem Rev* 116(5):2826–2885
- Chude-Okonkwo U, Malekian R, Maharaj B, Vasilakos A (2017) Molecular communication and nanonetwork for targeted drug delivery: a survey. *IEEE Commun Surv Tut* 19(4):3046–3096
- Felicetti L, Femminella M, Reali G, Lió P (2014) A molecular communication system in blood vessels for tumor detection. In: *ACM 1st annual international conference nanoscale computing and communication*, Atlanta
- Ghavami S, Lahouti F (2017) Abnormality detection in correlated Gaussian molecular nano-networks: design and analysis. *IEEE Trans NanoBiosci* 16(3):189–202
- Kay SM (1998) *Fundamentals of statistical signal processing*, vol 2. Springer/Prentice Hall PTR, Upper Saddle River
- Mai TC, Egan M, Duong TQ, Di Renzo M (2017) Event detection in molecular communication networks with anomalous diffusion. *IEEE Commun Lett* 21(6):1249–1252
- Mosayebi R, Jamali V, Ghoroghchian N, Schober R, Nasiri-Kenari M, Mehrabi M (2017) Cooperative abnormality detection via diffusive molecular communications. *IEEE Trans NanoBiosci* 16(8):828–842
- Nakano T, Okaie Y, Kobayashi S, Koujin T, Chan CH, Hsu YH, Obuchi T, Hara T, Hiraoka Y, Haraguchi T (2017) Performance evaluation of leader-follower-based mobile molecular communication networks for target detection applications. *IEEE Trans Commun* 65(2):663–676
- Okaie Y, Nakano T, Hara T, Nishio S (2016) *Target detection and tracking by bionanosensor networks*. Springer, Singapore
- Wuab L, Qu X (2015) Cancer biomarker detection: recent achievements and challenges. *Chem Soc Rev* 44(10):2963–2997

Molecular Propagation

- [Brownian Motion](#)

Molecular Receiver Decoding

- [Molecular Bit Detection](#)

Molecular Target Detection

- [Molecular Event Detection](#)

Molecular, Biological, and Multiscale Communications

- [Applications of Molecular Communication Systems](#)

Multicarrier Index Keying with Orthogonal Frequency Division Multiplexing

- [Index Modulation for OFDM](#)

Multicast Mobility

- [Mobility in Multicast](#)

Multi-channel Modulation

- [Principle of OFDM and Multi-carrier Modulations](#)

Multicore Architectures

- [Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks](#)

Multilayer Wireless Networks

- [Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks](#)

Multimedia Security

Leo Yu Zhang¹ and Kai Zeng²

¹School of Information Technology, Deakin University, Geelong, VIC, Australia

²George Mason University, Fairfax, VA, USA

Synonyms

[Digital right management](#)

Definitions

Multimedia is content composed of different types of digital objects such as text, audio, image, and video. Multimedia security addresses the protection of the intellectual property and the privacy of multimedia during acquisition, storage, processing, transmission, consumption, and disposal.

Historical Background

Advances in digital technologies have created significant changes in the way how multimedia is produced, distributed, marketed, and consumed. Typically, parties involved in multimedia business are *creator/owner*, *distributor*, and *consumer/user*. By preserving the security of multimedia, it is essentially protecting rights of all these parties. One notable milestone is the usage of Content Scramble System in the late 1990s, which is built on a stream cipher with 40-bit key, to protect commercially produced DVD discs by binding the content to a particular playback device (CSS 1996). It is later superseded by newer schemes such as the Content Protection for Recordable Media (CPR 2001) or the Advanced Access Content System (AAC 2005) due to low security strength. Similarly, the primary protection method for either the analog television or the digital television is encryption, such as VideoGuard (VG1 2012). The encryption system is built into the hardware of platform-supplied set-

top boxes so as to provide conditional access of the multimedia content. Additionally, it is common that the *creator* or *distributor* embeds a watermark-like logo in the broadcast content to increase brand recognition and assert ownership. The protection method used for online streaming television, such as the ones provided by Netflix and Hulu, is similar to dedicated terrestrial television, but this time the conditional access control is guaranteed by different technologies, such as the Common Encryption (CEN [2016](#)) and the Encrypted Media Extensions (EME) ([2017](#)). For example, the EME, recommended by W3C, provides copy protection of streaming content by providing a communication channel between web browsers and agent software.

All the abovementioned protection methods fit the centralized business model, where the *creator/distributor* plays the core role, but, driven by different technologies, e.g., cloud computing, IoT, healthcare personnel, and social networks, the business model itself keeps evolving. For example, the patient provides the health record to a hospital's server for personnel Medicare service, and people upload photos to Flickr and share them with friends. Under this decentralized scenario, there is no (strong) monetary incentive for *owners* to deploy a complex digital rights management (DRM) system, yet they still concern about privacy leakage. Despite the lack of direct monetary gain, the ubiquitous privacy concerns and users' unwillingness to participate continuously promote collaboration among legislative institution, industry, and academia to design new mechanisms to preserve the rights of different parties. For example, HIPAA ([HIP 1996](#)) regulates transactions of electronic healthcare data, Google's Encrypted BigQuery (EBQ [2015](#)) offers client-side encryption of user's query, and the prototype PIXEK (Pix [2018](#)) provides an end-to-end protection mechanism for their photos on the cloud.

Foundations

From the discussion above, the protection of multimedia, especially when the media object is with

low commercial value, requires the effort from law enforcement agency, coordination between different parties, and the appropriate usage of the technical measures. Here we mainly focus on technical measures, and the fundamental techniques used for multimedia security are discussed as follows.

Image/Video Source Identification

Image source identification aims to univocally link the image content to the device (i.e., camera) that captured it. Currently, the most promising technology to achieve this task exploits the detection of the sensor pattern noise (SPN) left by the acquisition device (Lukas et al. [2006](#)). This footprint is universal (i.e., every sensor introduces one) and unique (i.e., two SPNs are different even in case of images captured by cameras of the same brand and model). Machine learning techniques, such as support vector machine (SVM), are usually used to classify different image sources (cameras) according to the extracted features (fingerprints) of SPNs. For static images, SPN has been proven to be robust to common processing operations like JPEG compression (Lukas et al. [2006](#)). The photo response non-uniformity (PRNU) feature can be used for the case with severe compression (Milani et al. [2012](#)). However, compression and distortion in digital videos can usually increase the false-alarm probability. In addition, when considering video streaming over wireless networks, blocking and blurring effects caused by packet loss should be tackled for effective camera source identification. Detailed discussions can be found in Chen et al. ([2015](#)).

Digital Watermarking

Digital watermarking relies on the fact that multimedia data is noise-tolerant. It works by first imperceptibly embedding a unique watermark (fingerprint) into each copy of the considered multimedia content and then detecting the existence of the unique watermark from a suspicious copy. The research focus for watermarking is to investigate the trade-offs among embedding capacity, imperceptibility, and robustness. Most, if not all, digital watermarking systems are devel-

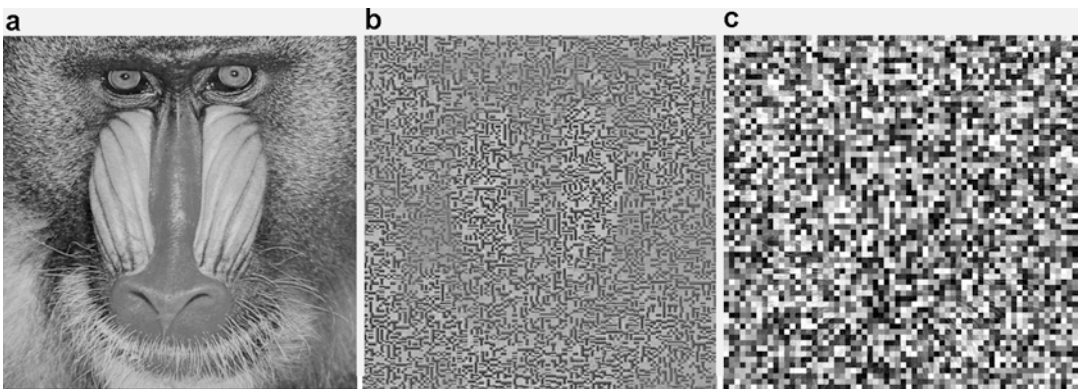
oped from the following two seminal designs: the spread spectrum system (Cox et al. 1997) and the quantization index modulation system (Chen and Wornell 2001). Detailed discussion about different digital watermarking systems can be found in Cox et al. (2007) and Barni and Bartolini (2004).

Format Compatible Encryption

Format Compatible Encryption (FCE) relies on the fact that multimedia data is always encoded and then stored under certain formats, such as WAV or MP3 for speech, JPEG or PNG for image, and WAV or MP4 for video. By modifying the behavior of encoder, FCE aims to make the syntax of encrypted data compliant with a standard decoder, i.e., it is decodable without decryption but suffered by (controllable) quality degradation. For authenticated users with the key, the joint encryption and encoding process can be removed without quality loss. For example, the JPEG compression of still images is composed by the DCT transformation, quantization, and Huffman coding processes. A FCE scheme for JPEG could be secretly shuffling or quantizing the transformation coefficients, as shown in Fig. 1. Detailed review of FCE scheme for multimedia data can be found in Stutz and Uhl (2012) and Massoudi et al. (2008).

Privacy-Aware Processing

Privacy-aware processing of multimedia data could rely on two different strategies: the one based on trusted execution environment and the one based on secure Multi-Party Computation (MPC). For the first class, representative technology includes Intel Software Guard Extensions (SGX) and AMD secure encrypted virtualization. For example, the work in Ohrimenko et al. (2016) presents a design that allows multiple hospitals to perform collaborative data analytics while guaranteeing the privacy of their patient datasets by using SGX. The MPC-based privacy-aware processing technology stems from the simple fact that all arithmetic operation can be expressed as a function of Boolean circuits, while Boolean circuit itself has a secured counterpart called Yao's Garbled circuit (Yao 1986). For example, Garbled circuit is used for server-side privacy-preserving image denoising (Zheng et al. 2017). Moreover, it is generally acknowledged that employing Garbled circuit, homomorphic encryption, and secret sharing to devise hybrid privacy-aware protocol for specific tasks will reduce cost and thus bring computation and bandwidth gain. Details about Intel Software Guard Extensions can be found in SGX (2015), and a MPC technique can be found in Lindell (2003).



Multimedia Security, Fig. 1 A FCE example on 512×512 image "Baboon": (a) Original image. (b) Standard JPEG decoding result when the 8 most significant DCT

coefficients are randomly permuted. (c) Standard JPEG decoding result when all the DC coefficients are encrypted

Key Applications

Multimedia security technology is a key-enabler for many businesses and services. The applications described below cover its typical usage.

- **Video Conferencing:** Video conferencing is a special way to conduct meetings; it allows users to communicate in real-time remotely so as to save a ton in travel cost. It is widely used to enable discussion on the most important perspectives of business right up to executive level. Possible security threats to video conferencing could be enormous, but the typical way to harden the security is to combine an identity-based access control policy and data encryption by deploying dedicated video conferencing systems, such as Cisco Webex or Skype for Business.
- **Online Streaming:** It is reported that Netflix, after hitting 50.85 million subscribers in 2016, has surpassed the total number of subscribers of all US cable companies. By shifting from cabled TV to streaming service, users are not restricted to special playback devices and able to enjoy content anytime anywhere. The secure multimedia streaming is commonly achieved by two mechanisms: using encryption to enforce conditional access control and using watermarking to deter pirates distribution. Beside technical measures, law enforcement has been advocated by *creators/distributors* under this centralized business model. For example, in a recent response to a call from NTIA, MPAA suggests criminal charges against online video piracy and coordination between a broad range of online service providers (MPA 2018).
- **Privacy-Preserving Multimedia Processing:** Hosting or contributing personal data to a public server for personalized service becomes popular due to the maturity of artificial intelligence and cloud computing. In this case, content *owners* are likely to be individuals and the value of a single piece of data is not high but the whole dataset is the key to success of

many businesses. For example, recommender systems collect and analyze user's past behavior and decisions to generate recommendations. Since privacy leakage could happen at any stage of the life cycle of multimedia data, it essentially requires imposing protect to data acquisition, storage, processing, transmission, consumption, and disposal.

Cross-References

- [Access Control](#)
- [Data-Driven Security](#)
- [IoT Security](#)
- [Mobile Security and Privacy](#)

References

- (1996) Content Scramble System. <http://www.dvcca.org/css.aspx>. Accessed 02 Aug 2018
- (1996) Summary of the HIPAA Security Rule. <https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html>. Accessed 02 Aug 2018
- (2001) 4C Entity CPRM Specification. <http://www.4centity.com/specification.aspx>. Accessed 02 Aug 2018
- (2005) Advanced Access Content System Specification. <https://www.aacsla.com/specifications/>. Accessed 02 Aug 2018
- (2012) VideoGuard DRM Design Guides. <https://www.cisco.com/c/en/us/support/video/video-guard-drm/products-implementation-design-guides-list.html>. Accessed 02 Aug 2018
- (2015) Encrypted BigQuery Client. <https://opensource.google.com/projects/encrypted-bigquery-client>. Accessed 02 Aug 2018
- (2015) Intel Software Guard Extensions. <https://software.intel.com/en-us/sgx>. Accessed 02 Aug 2018
- (2016) ISO Common Encryption Protection Scheme. <https://www.w3.org/TR/eme-stream-mp4/>. Accessed 02 Aug 2018
- (2017) W3C Recommendation Encrypted Media Extensions. <https://www.w3.org/TR/encrypted-media/>. Accessed 02 Aug 2018
- (2018) Pixek. <https://pixek.io/>. Accessed 02 Aug 2018
- (2018) Response of the Motion Picture Association of America to National Telecommunications and Information Administration Internet Priorities Inquiry. <https://www.mpaa.org/wp-content/uploads/2018/07>

/180717-MPAA-response-to-NTIA-internet-priorities-inquiry.pdf. Accessed 02 Aug 2018

- Barni M, Bartolini F (2004) Watermarking systems engineering: enabling digital assets security and other applications. CRC Press, Boca Raton, FL
- Chen B, Wornell GW (2001) Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Trans Inf Theory* 47(4):1423–1443
- Chen S, Pande A, Zeng K, Mohapatra P (2015) Live video forensics: source identification in lossy wireless networks. *IEEE Trans Inf Forensics Secur* 10(1):28–39
- Cox I, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. *IEEE Trans Image Process* 6(12):1673–1687
- Cox I, Miller M, Bloom J, Fridrich J, Kalker T (2007) Digital watermarking and steganography. Morgan Kaufmann, San Francisco, CA
- Lindell Y (2003) Composition of secure multi-party protocols: a comprehensive study. Springer Berlin Heidelberg, New York
- Lukas J, Fridrich J, Goljan M (2006) Digital camera identification from sensor pattern noise. *IEEE Trans Inf Forensics Secur* 1(2):205–214. <https://doi.org/10.1109/TIFS.2006.873602>
- Massoudi A, Lefebvre F, De Vleeschouwer C, Macq B, Quisquater J (2008) Overview on selective encryption of image and video: challenges and perspectives. *Eurasip J Inf Secur* 5:1–18
- Milani S, Fontani M, Bestagini P, Barni M, Piva A, Tagliasacchi M, Tubaro S (2012) An overview on video forensics. *APSIPA Trans Signal Inf Process* 1: 1–18
- Ohrimenko O, Schuster F, Fournet C, Mehta A, Nowozin S, Vaswani K, Costa M (2016) Oblivious multi-party machine learning on trusted processors. In: *USENIX security symposium*, pp 619–636
- Stutz T, Uhl A (2012) A survey of H.264 AVC/SVC encryption. *IEEE Trans Circuits Syst Video Technol* 22(3):325–339
- Yao ACC (1986) How to generate and exchange secrets. In: *27th annual symposium on foundations of computer science (FOCS)*, pp 162–167
- Zheng Y, Cui H, Wang C, Zhou J (2017) Privacy-preserving image denoising from external cloud databases. *IEEE Trans Inf Forensics Secur* 12(6):1285–1298

Multiparty Key Agreement for Wireless Networks

- [Group Key Agreement for Wireless Networks](#)

Multipath Transport Protocol

- [Collaborative Multipath Transmission](#)

Multiple Access in High Frequency

- [Millimeter Wave NOMA](#)

Multiple Access Methods

- [Multiple Access Techniques](#)

Multiple Access Technique for Cellular Wireless Networks

Qi-Yue Yu and Hong-Chi Lin
School of Electronics and Information
Engineering, Harbin Institute of Technology,
Harbin, Hei Longjiang, China

Synonyms

[Division multiplexing](#)

Definitions

Multiple access is the technique that multiple signals simultaneously access the same terminal, i.e., a base station, or one terminal simultaneously transmit multiple signals to different user terminals. Different resources, i.e., time, frequency, code, etc., can be assigned to different signals to avoid or reduce the multiple access interferences (MAC).

History Background

During the last decades, multiple access technique has been widely used in mobile communications and provided increasable capability and various services, i.e., from traditional voice to multimedia communications, as shown in Fig. 1. It plays a fundamental and important role in cellular networks.

A cellular always includes one base station (BS) and many users that are located in its coverage. And the coverage area of one cellular is generally determined by the transmit power of its BS and mathematically expressed as a hexagon, as shown in Fig. 2. There are two basic conceptions, one is uplink transmission and the other is downlink transmission. Uplink transmission means that many users transmit their signals to the BS, and downlink is the opposite processing that the BS simultaneously transmits signals to the multiple users. And multiple access is a technique that is used to reflect both uplink and downlink access processing.

The multiple access technology is always viewed as one of the most important techniques for cellular wireless communications. With the

increasing of data rates, the multiple access technology updates correspondingly.

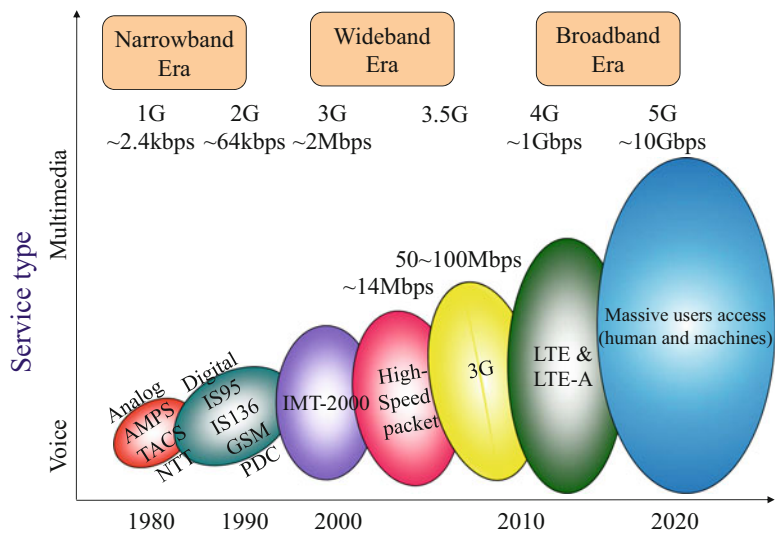
There are lots of works that have been done on this topic, from the first-generation (1G) frequency-division multiple access (FDMA), the second-generation (2G) time-division multiple access (TDMA), the third-generation (3G) code-division multiple access (CDMA) (Adachi et al. 2005; Schulze and Luders 2005) to the fourth-generation (4G) orthogonal frequency-division multiple access (OFDMA). The evolution of multiple access techniques has pushed for a rapid advancement of wireless communications. Recently, many researchers believe that, for the coming fifth generation (5G), it will be non-orthogonal multiple access (NOMA) era.

Classification

FDMA

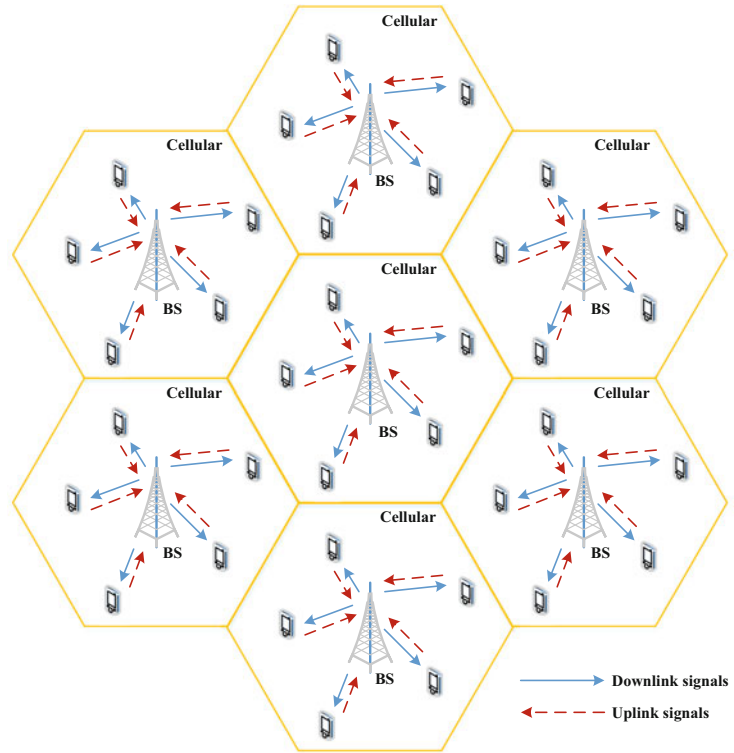
Frequency-division multiple access (FDMA), as its name, utilizes frequencies to divide different users, and each user is allocated a carrier frequency. It is noted that an extra frequency band

Multiple Access Technique for Cellular Wireless Networks, Fig. 1
The development of wireless communications



Multiple Access Technique for Cellular Wireless Networks, Fig. 2

A diagram of cellular networks



is necessary between any adjacent frequencies, to reduce the sidelobes' interferences. For downlink, different users occupy different frequencies; therefore the BS can realize simultaneous transmission. For uplink, there is no multiple access interference (MAI) among different users, since each user occupies its respective frequency. A diagram of FDMA is shown in Fig. 3a.

There are two major advantages of FDMA, one is that it is easy to realize and the other is there is no MAI. The disadvantages are also obvious, for example, it occupies a wider frequency bandwidth, leading to a lower spectrum efficiency (SE). It is interesting to design filters to reduce sidelobes, i.e., filter bank. If there is only a few users and there are enough frequency resources, FDMA is a promising technique.

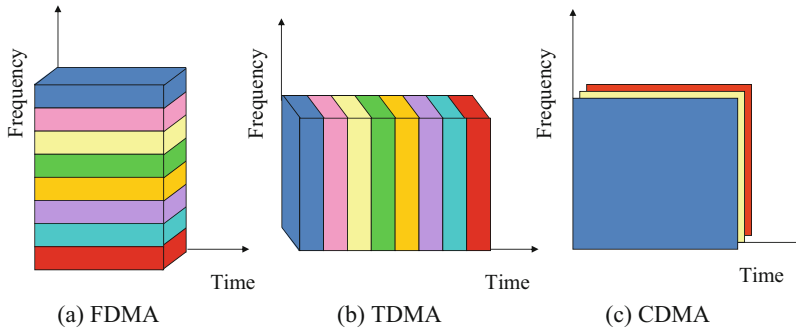
TDMA

Time-division multiple access (TDMA) is similar to the FDMA case, which explores time instead of frequency to distinct different users, as shown in Fig. 3b. The major difference between TDMA

and FDMA is that the synchronization is very important for TDMA. Without accurate time synchronization, there will be serious MAI among users, and even the performances of the whole system can be destroyed. Thus, many works have been done on the synchronization problem of TDMA.

CDMA

Code-division multiple access (CDMA) is widely used in 3G system, in which every user occupies the same time and frequency, and users are separated by different spreading sequences/vectors, as shown in Fig. 3c. Each user is assigned a spreading sequence/vector, and these spreading vectors are expected to be orthogonal to each other. The classical structure is named as direct-sequence (DS)-CDMA; besides this basically one, there are generally two approaches: single-carrier (SC)-CDMA and multicarrier (MC)-CDMA, as shown in Figs. 4 and 5, respectively (Adachi et al. 2005). Both SC- and MC-CDMA can flexibly provide variable-rate transmissions, with retaining multi-



Multiple Access Technique for Cellular Wireless Networks, Fig. 3 A diagram to show the features of FDMA, TDMA, and CDMA

ple access capability. Actually, OFDM is a special case of MC-CDMA when the spreading factor is one.

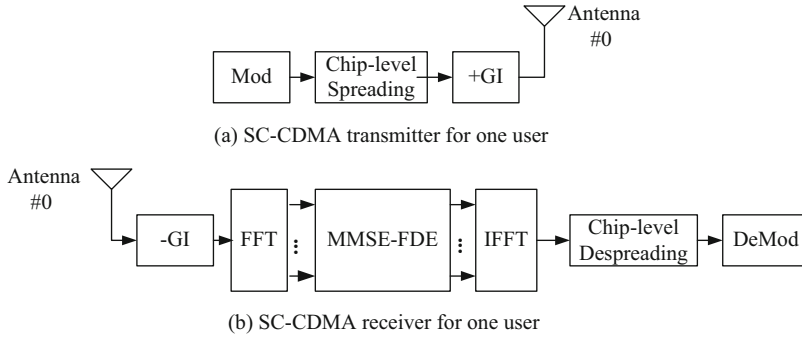
The SC-CDMA and MC-CDMA, respectively, explore time-domain and frequency-domain spreading (chip-level spreading is used for further discussion). At the transmitter, after the binary information data is modulated, the data-modulated symbols are then spread, time domain spread for SC-CDMA, and frequency domain spread for MC-CDMA. Without/with IFFT processing for SC-/MC-CDMA, the processed signals insert GI and then transmit to the channels.

At the receiver, frequency-domain equalization (FDE) can be used to improve the BER performances. It has been proved that the use of minimum mean square error (MMSE) weight can provide the best BER performances among various FDE weights, since the MMSE weight can provide the best compromise between the noise enhancement and suppression of frequency-selective fading channels. Simulation results show that MC-CDMA with MMSE-FDE provides much better BER performances than SC-CDMA with traditional coherent rake combining. It has been derived that SC-CDMA with proper FDE, instead of the traditional rake combining, can achieve good BER performances that are comparable to MC-CDMA.

For uplink CDMA systems, different users' signals are asynchronously arrived at the BS via different fading channels, leading to seri-

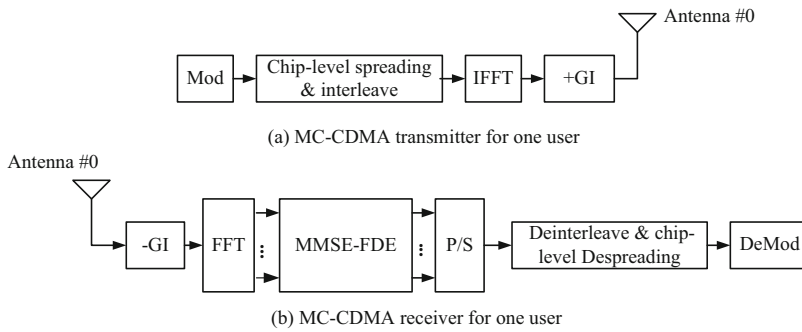
ous MAIs, which limit the uplink capacity. The capability of MAI suppression achievable with pure CDMA is not sufficient. It is an important research topic to suppress the effect of MAI. There are many interesting methods, such as multiuser detection (MUD), soft-iterative multistage receiver, etc.

In papers (Liu and Adachi 2006; Yu et al. 2013, 2014), two-dimensional (2D) block spread CDMA is proposed, which can be applied to solve the MAI and achieve frequency diversity gain in frequency-selective fading channels. The 2D block spread CDMA consisted of chip-level and block-level spreading, and they are implemented as different roles. The chip-level spreading plays the same part as traditional SC-/MC-CDMA, which can achieve frequency diversity gain by using MMSE-FDE in the receiver. In 2D block spread CDMA, block-level spreading is performed to each block after chip-level spreading. Before the transmission, the GI, which is larger than or equal to the maximum delay among different users, is inserted. At the receiver, after removing the GI, block-level de-spreading is performed in order to remove the MAI. If the maximum timing offset among users is within the GI length, perfect removal of MAI is possible in the case of block fading, i.e., the channel stays constant during at least one block. Both chip-level spreading codes and block-level spreading codes can be constructed using the orthogonal variable spreading factor code tree. The 2D block SC-CDMA and MC-CDMA are given in Figs. 6 and 7.



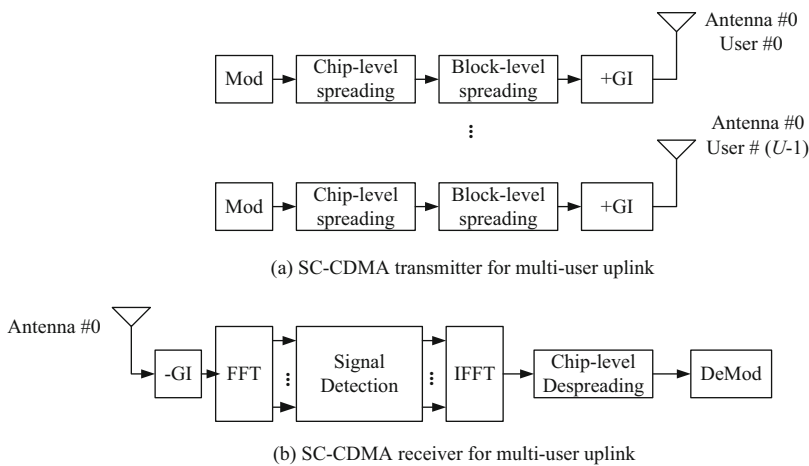
Multiple Access Technique for Cellular Wireless Networks, Fig. 4 A diagram of SC-CDMA transmitter and receiver, where “Mod” and “DeMod” indicate modula-

tion and demodulation, respectively, and “FDE” means frequency-domain equalization



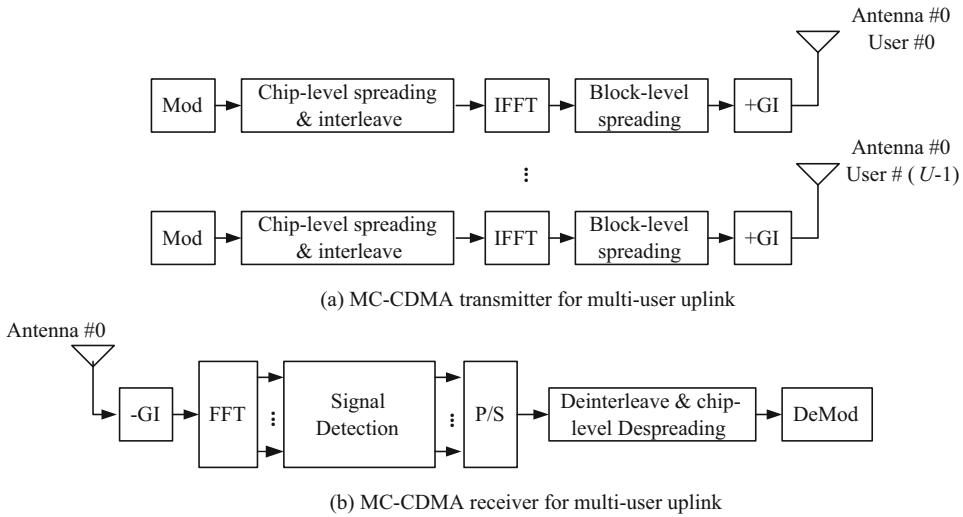
Multiple Access Technique for Cellular Wireless Networks, Fig. 5 A diagram of MC-CDMA transmitter and receiver, where “Mod” and “DeMod” indicate modula-

tion and demodulation, respectively, and “FDE” means frequency-domain equalization



Multiple Access Technique for Cellular Wireless Networks, Fig. 6 A diagram of 2D block SC-CDMA transmitter and receiver, where “Mod” and “DeMod” indicate

modulation and demodulation, respectively, and “FDE” means frequency-domain equalization



Multiple Access Technique for Cellular Wireless Networks, Fig. 7 A diagram of 2D block MC-CDMA transmitter and receiver, where “Mod” and “DeMod” indicate

modulation and demodulation, respectively, and “FDE” means frequency-domain equalization

IDMA

Interleave-division multiple access (IDMA) is also an interesting multiple access scheme, which was proposed in Ping Li et al. (2006) by Prof. Li, as shown in Fig. 8. Actually, IDMA and CDMA work in a similar way, both of which occupy all the same time- and frequency-domain to transmit data information. There are several differences between IDMA and CDMA, such as: (1) in CDMA, transmit symbols are modulated by different spreading sequences/vectors to distinguish different users, whereas IDMA exploits different interleavers to separate different signals, and (2) at a receiver, de-spreading is used to recover transmitted signals in CDMA, while iterative algorithm, just like decoding algorithms for low-density parity-check (LDPC), is used in IDMA. IDMA can be flexibly applied in different modulations and channel conditions.

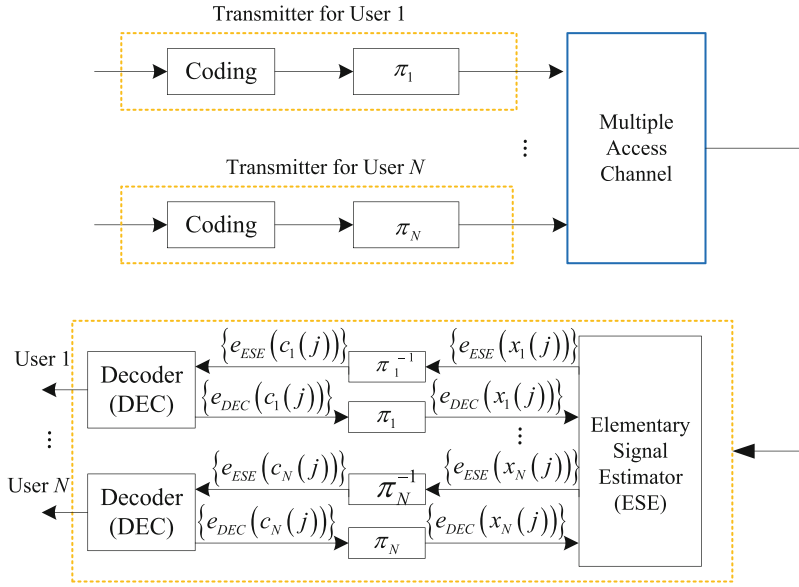
OFDMA

Orthogonal frequency-division multiple access (OFDMA) or orthogonal frequency-division multiplexing (OFDM) is one of the most appreciated multiple access schemes. Actually, OFDM has been widely used in 4G; WLAN, i.e., IEEE 802.11a/g; and digital audio and video broadcasting (DAB, DVB) systems. It also uses orthog-

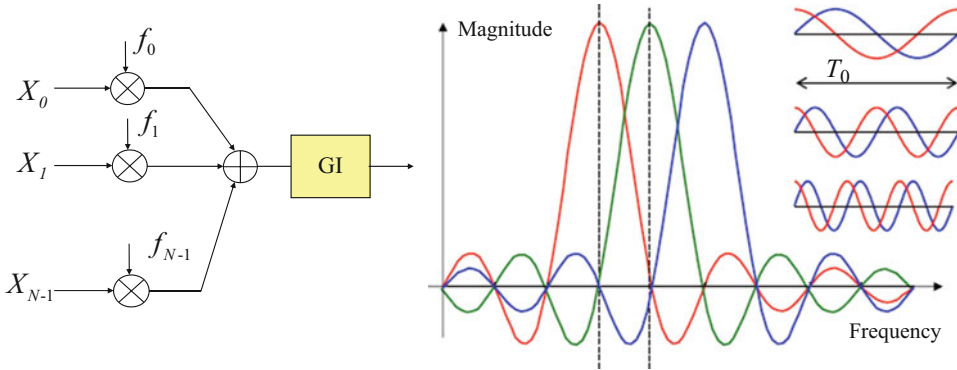
onal subcarriers to separate users; however its bandwidth is much smaller than the FDMA case. And it doesn't need extra bandwidth to avoid the effect of sidelobe. Generally speaking, a subcarrier's spectrum of an OFDM symbol is the Fourier transform of a rectangular window of symbol duration T_s , which is a sinc-function with zeros at $\frac{1}{T_s}$. And the centers of other carriers are put in these zeros; thereby, the subcarriers are orthogonal to each other, as shown in Fig. 9.

DAB always explores COFDM (coded OFDM), in which the data transmitted on the subcarriers is protected by forward error correction (FEC) coding. COFDM also allows different groups of bits to be protected with a different strength code rate. Considering the subcarriers are subject to flat fading, interleaving can be used to improve the bit error rate (BER) performance of OFDM. For an OFDM system, interleaving is generally classified as time and frequency interleaving or random and regular interleaving.

An OFDM modulator mainly consists of two parts: IDFT (inverse discrete Fourier transform) and guard interval (GI) that is also called cyclic prefix (CP), as shown in Fig. 10. The IDFT makes OFDM practically feasible. And we don't need many filters and oscillators any more. In reality,



Multiple Access Technique for Cellular Wireless Networks, Fig. 8 A diagram of IDMA transmitter and receiver given in Ping Li et al. (2006)



$$\text{where } x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, n = 0, 1, \dots, N-1$$

Multiple Access Technique for Cellular Wireless Networks, Fig. 9 A diagram to show the spectrum of OFDM signals

the actual processing utilizes IFFT (inverse fast Fourier transform) instead of IDFT.

The GI (or CP) should be larger than, or at least equal to, the maximum multipath delay. Hence, it can avoid the inter-symbol interferences (ISI) caused by the multipath delay. The GI can efficiently avoid the effect of ISI; however, it will decrease the throughput of an OFDM

system. Therefore, some papers (Liu et al. 2017a) try to remove GI, with acceptable detection complexity.

There are lots of papers focusing on OFDM system, from synchronization, nonconstant power envelope, to the channel estimation, etc. For the synchronization issue, the frequency-domain nature of OFDM makes the effect of

several synchronization errors to be explained with the properties of DFT (discrete Fourier transform). In fact, OFDM synchronization functions can be performed either in time or frequency domain, which is different from the traditional single-carrier systems. To decrease the effect of asynchronization, low sidelobes of OFDM symbols are appreciated. Filtered multitone (FMT) modulation is a useful scheme, which is a kind of filter bank implementation of multicarrier system (Proakis 2009).

Another major problem of OFDM is relatively high PAPR (peak-to-average power ratio) issue. In general, the large signal peak occurs when the signals in the sub-channels add constructively in phase. Such large signal peaks may result in clipping of the signal voltage in a DAC (digital-to-analogue converter), when the OFDM signal is synthesized digitally. It may saturate the power amplifier, leading to the intermodulation distortion. To avoid the PAPR problem, there are generally two kinds of methods. One is exploring distortion methods, i.e., clipping, filtering, peak cancellation, etc.; and the other is exploring the methods without distortion, for example, selective mapping, partial transmit sequence, etc (Proakis 2009).

Figure 10 shows the implementation of an OFDM symbol. After passing through IFFT processor, the values are fed to DACs. And the low-pass filtered signals of the real and imaginary streams are amplitude modulated onto RF carrier and then added together and sent through a band-pass filter (BPF) to the transmit antenna.

NOMA

Non-orthogonal multiple access (NOMA) explores the same resource block to serve different users, thus transfers information of multiple users superposition (Ding et al. 2017; Dai et al. 2015). It is different from orthogonal multiple access (OMA), which assigns orthogonal resources to different users. Owing to the orthogonal feature, different signals can be separated without MAI. However, the spectrum efficiency (SE) of OMA schemes is low.

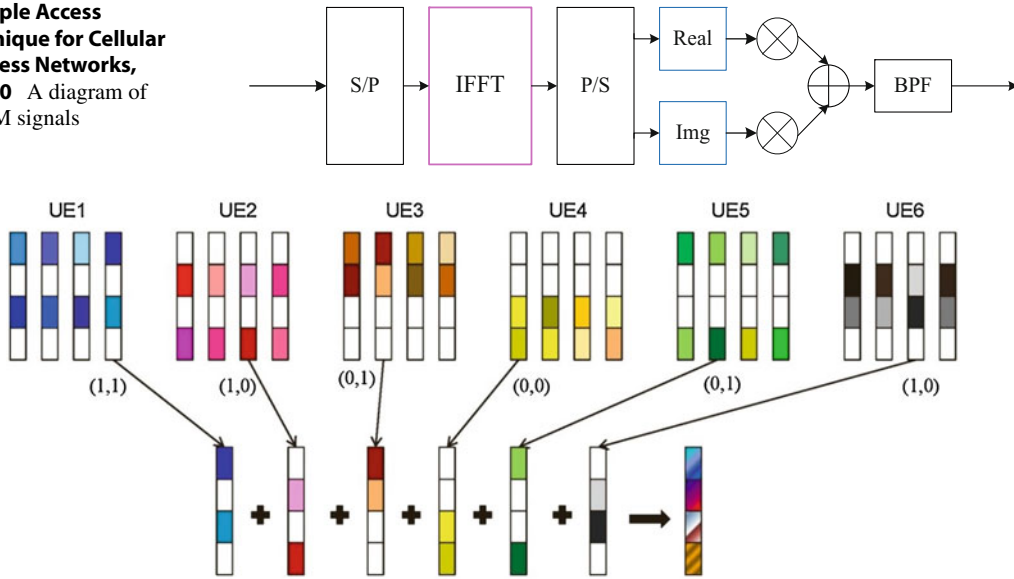
NOMA can improve SE by overloaded information, which can be realized in power domain, by different modulation amplitudes, and/or code domain, via multidimensional constellation codebooks. Therefore, the core of all the emerging multiple access techniques is to optimally combine modulation and coding techniques to support high spectrum efficiency.

Nowadays, the multiple access schemes of 5G need to support massive users of both human and machines. High SE multiple access schemes become emergency. Many novel multiple access techniques have been proposed, such as power-domain NOMA, sparse code multiple access (SCMA), pattern division multiple access (PDMA), multiuser shared access (MUSA), etc. Compared to the well-known OMA schemes such as OFDMA, NOMA provides improved SE by transmitting overloaded information in the same time resources or frequency resources (Ding et al. 2017; Dai et al. 2015).

Power-domain NOMA allocates different power to different users according to their different channel state information (CSI). In general, a user with better CSI is assigned a small power; in contrast, a user with worse CSI is assigned a large power to keep fairness. And successive interference cancelation (SIC) algorithm is utilized at the receiver to decode the superimposed signal in sequence. Power-domain NOMA is easy to realize with low complexity, and it is applicable for Internet of Things networks.

SCMA, which was first proposed in 2013 (Nikopour and Baligh 2013), can also be viewed as a type of NOMA technique, which works based on traditional DS-CDMA technique, as shown in Fig. 11. Its input binary bits first pass through a modulator to map the binary bits to complex symbols, and then direct sequence spreading is performed to all mapped complex symbols. Compared to the DS-CDMA, SCMA combines modulation and spreading, and thus it can achieve shaping gain of multidimensional constellations due to its optimal SCMA codebook design (Nikopour and Baligh 2013; Nikopour et al. 2014; Zhou et al. 2017).

Multiple Access Technique for Cellular Wireless Networks, Fig. 10 A diagram of OFDM signals



Multiple Access Technique for Cellular Wireless Networks, Fig. 11 A diagram of SCMA system that serves six users with four resources

Key Applications

Cellular systems, Satellite communications, Internet of Things (IoT) systems, etc.

Cross-References

- [Beam Division Multiple Access \(BDMA\) Transmission](#)
- [Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation](#)
- [Non-orthogonal Multiple Access \(NOMA\)](#)
- [Multiple Access Techniques](#)
- [Sparse Code Multiple Access](#)

References

- Adachi F, Garg D, Takaoka S, Takeda K (2005) Broadband CDMA techniques. *IEEE Wirel Commun* 12(2):8–18
- Dai L, Wang B, Yuan Y, Han S, Chih-Lin I, Wang Z (2015) Nonorthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun Mag* 53(9):74–81
- Ding Z, Lei X, Karagiannidis GK, Schober R, Yuan J, Bhargava VK (2017) A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. *IEEE J Sel Areas Commun* 35(10):2181–2195
- Liu X, Chen H, Chen S, Meng W (2017a) Symbol cyclic shift equalization algorithm – a CP-free OFDM/OFDMA system design. *IEEE Trans Veh Technol* 66(1):282–294
- Liu X, Chen H, Lyu B, Meng W (2017b) Symbol cyclic shift equalization PAM-OFDM – a low complexity CP-free OFDM scheme. *IEEE Trans Veh Technol* 66(7):5933–5946
- Liu L, Adachi F (2006) 2-Dimensional OVFSF spread/chip-interleaved CDMA. *IEICE Trans Commun* E89-B(12):3363–3375
- Nikopour H, Baligh H (2013) Sparse code multiple access. In: *Proceedings of the IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pp 332–336
- Nikopour H, Yi E, Bayesteh A, Au K, Hawryluck M, Baligh H, Ma J (2014) SCMA for downlink multiple access of 5G wireless networks. In: *Proceedings of the IEEE GLOBECOM*, pp 3940–3945
- Ping L, Liu L, Wu K, Leung WK (2006) Interleave division multiple-access. *IEEE Trans Wirel Commun* 5(4):938–947
- Proakis JG (2009) *Digital communications*, 5th edn. Publishing House of Electronics Industry, Beijing
- Schulze H, Luders CN (2005) *Theory and applications of OFDM and CDMA: wideband wireless communications*. John Wiley & Sons, pp 269–272
- Yu Q, Meng W, Adachi F (2013) Two-dimensional block-spread CDMA relay using virtual-four-antenna STCDD. *IEEE Trans Veh Technol* 62(8):3813–3827

Yu Q, Meng W, Adachi F (2014) Combined code reuse scheme with two-dimensional OVFS codes assignment algorithm for uplink multi-user/multi-rate block spread multi-cellular CDMA. *Wirel Commun Mob Comput* 14(13):1314–1328

Zhou Y, Yu Q, Meng W, Li C (2017) SCMA codebook design based on constellation rotation. *IEEE Int Conf Commun* 1–6

Multiple Access Techniques

Fan Jiang¹, Zijun Gong¹, Kun Hao^{2,3}, and Yan Zhang^{2,3}

¹Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

²Department of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada

³Tianjin Chengjian University, Tianjin, China

Synonyms

[Channel access methods](#); [Channel access techniques](#); [Multiple access methods](#)

Definition

Multiple access techniques allow multiple terminals to share the common communication medium based on multiplexing. The multiplexing is provided in the physical layer, and the shared communication medium can be wired cables or wireless spectrum. The availability of the communication medium is limited, for example, limited frequency band and limited time. Therefore, multiple access techniques are essential to maintain successful communication among multiple devices. The basic well-known multiple access techniques are frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and spatial division multiple access (SDMA). A class of non-orthogonal multiple

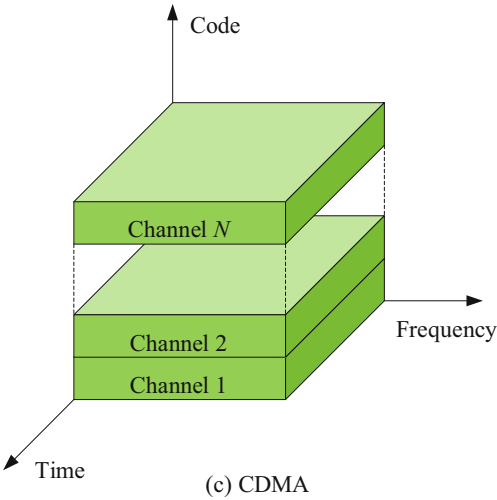
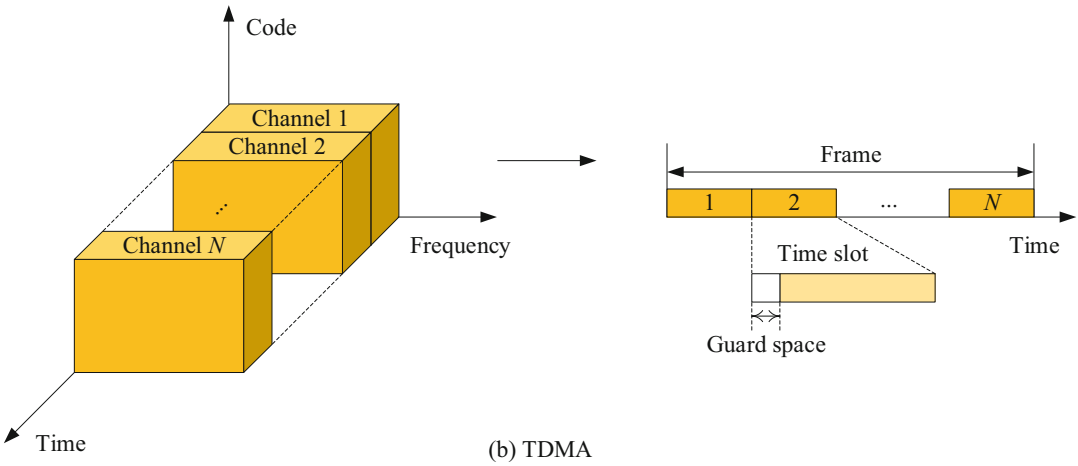
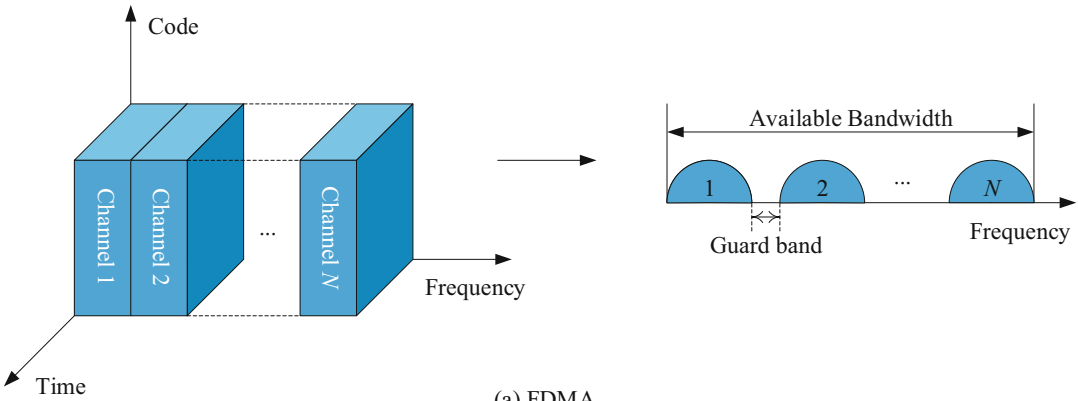
access (NOMA) techniques has recently gained much interest. The control mechanism to distribute channels to users is also known as medium access control (MAC).

Historical Background

The applications of the multiple access techniques present in every generation of wireless communication systems. In the first US analog cellular system, AMPS (Advanced Mobile Phone System) is based on FDMA. TDMA and CDMA dominate in the second-generation (2G) wireless communication systems. For example, in GSM (Global System for Mobile communication), TDMA is adopted for distinguished users, while the CDMA is adopted in IS-95 (Interim Standard 95) systems. The varieties of multiple access techniques based on TDMA and CDMA have been widely adopted in ad hoc sensor networks. CDMA becomes the dominant multiple access technique in the third-generation (3G) wireless communication systems, with the prevailing worldwide standards known as CDMA2000, WCDMA (Wideband CDMA), and TD-SCDMA (Time Division – Synchronous CDMA). In fourth generation (4G), the hybrid multiple access techniques, including orthogonal FDMA (OFDMA), TDMA, and SDMA, are preferred in multiple-input multiple-output (MIMO) systems (Miao et al. 2016). With the increase of the MIMO size, massive MIMO, also known as large-scale MIMO, together with millimeter wave communications, becomes one of the promising techniques in fifth-generation (5G) wireless communication systems (Marzetta 2010; Dai et al. 2015; Sun et al. 2015). In massive MIMO, the SDMA techniques, for example, BDMA (beam division multiple access), are recently studied in Sun et al. (2015).

Foundations

The standard FDMA, TDMA, and CDMA techniques can be further illustrated in Fig. 1. As it is shown in Fig. 1a, when FDMA is adopted,

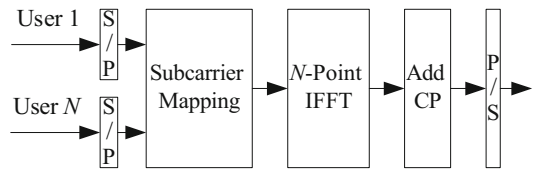


Multiple Access Techniques, Fig. 1 Illustration of FDMA/TDMA/CDMA techniques

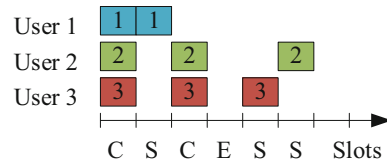
the system bandwidth is divided into several narrowbands, with nonoverlapping frequency slots, (i.e., the adjacent frequency bands are separated by a guard band). Usually, in a FDMA system, each user occupies a pair of frequency bands (i.e., frequency division duplexing (FDD)). One is for forward channel (downlink), and the other is for the reverse channel (uplink). FDMA was the initial multiple access technique for cellular systems. An advanced form of FDMA is the orthogonal frequency division multiple access (OFDMA), which is shown in Fig. 2. In OFDMA, each user is assigned a subset of subcarriers, and the low-complexity fast Fourier transform (FFT) operations can be used at both the transmitter and receiver side in the system (Miao et al. 2016).

In TDMA systems, the data frame is segmented into nonoverlapping time slots, separated by guard space. When the time slot is assigned to a user, the user takes up the entire frequency band for data transfer. An advanced form of TDMA is dynamic TDMA (DTDMA), where different time slots may be assigned to the same user according to the scheduling algorithm. For example, the user terminal may use the first time slot for the first data frame but use another time slot for the next data frame. DTDMA usually works with MAC protocol to randomly access the shared channel. Common examples for DTDMA include the slotted ALOHA in GSM systems (initial access stage for users) and carrier sense multiple access with collision avoidance (CSMA/CA) in IEEE 802.11. An example of the slotted ALOHA can be illustrated in Fig. 3 (Tanenbaum 2003). In slotted ALOHA, the time frame is divided into equal size slots. The data packet from each user is transmitted at the beginning of next slot. If collision occurs, retransmitted packet will select future slots with probability p , until successful. Specifically, in Fig. 3, collisions occur in the first and third slot, while packets 1, 2, and 3 from each user are successfully delivered in the second, sixth, and fifth slot, respectively. DTDMA schemes have been widely in ad hoc networks since the channel can be randomly accessed.

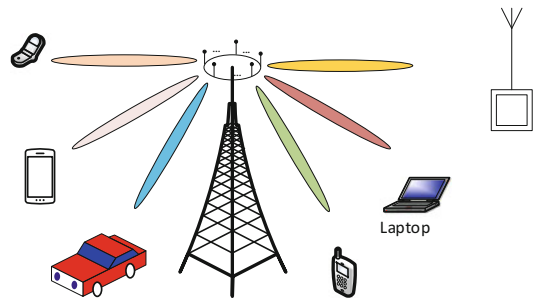
The CDMA scheme is based on spread-spectrum code. Signals from different users are



Multiple Access Techniques, Fig. 2 OFDMA block diagram



Multiple Access Techniques, Fig. 3 Slotted ALOHA, C collision, S successful, E empty



Multiple Access Techniques, Fig. 4 Illustration of BDMA

transferred simultaneously over the same carrier frequency. At the receiver side, the user signal is separated by cross-correlation of the received signal with each possible code sequence. Two typical forms of CDMA are the direct sequence spread spectrum (DSSS) and the frequency-hopping spread spectrum (FHSS) (Miao et al. 2016). The CDMA scheme is known to have low signal-to-noise ratio (SNR) working region, which indicates lower transmission power is required in the system. The near-far problem is a serious one in CDMA systems.

SDMA utilizes the spatial separation of the users to fully use the frequency spectrum. In Fig. 4, an example of SDMA techniques, i.e., BDMA is illustrated (Sun et al. 2015). As it is shown in Fig. 4, multiple beams are generated by using the phased array antennas. SDMA has been

one of the promising multiple access techniques for 5G.

Besides the orthogonal multiple access techniques, a class of NOMA schemes is also discussed for 5G (Dai et al. 2015). NOMA accommodates multiple users through non-orthogonal resource allocation. Generally, NOMA can be implemented via power or code domain multiplexing. Examples of NOMA schemes are power-domain NOMA, multiple access with low-density spreading code, sparse code multiple access, multiuser shared access, and so on (Dai et al. 2015).

Key Applications

Multiple access techniques are fundamental in wireless communication systems, including cellular networks, ad hoc networks, and vehicular networks, for multiple user terminals to access the shared common medium.

Cross-References

- [Non-orthogonal Multiple Access](#)
- [Massive MIMO](#)
- [Medium Access Control](#)
- [Resource Allocation](#)

References

- Dai L, Wang B, Yuan Y, Han S, I C, Wang Z (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities and future research trends. *IEEE Commun Mag* 53(9):74–81
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of Base Station antennas. *IEEE Trans Wirel Commun* 9(11):3590–3600
- Miao G, Zander J, Sung KW, Slimance B (2016) Fundamentals of mobile data networks. Cambridge University Press, New York
- Sun C, Gao XQ, Jin S, Matthaiou M, Ding Z, Xiao CS (2015) Beam division multiple access transmission for massive MIMO communications. *IEEE Trans Commun* 63(6):2170–2184
- Tanenbaum AS (2003) Computer networks. Prentice Hall PTR, Upper Saddle River

Multiple Network Paths

- [Collaborative Multipath Transmission](#)

Multi-provider Pricing

- [Oligopoly Pricing in Wireless Networks](#)

Multi-tone Modulation

- [Principle of OFDM and Multi-carrier Modulations](#)

Multuser Information Theory

- [Network Information Theory](#)

N

Named Data Networking

- [Information-Centric Networking: Basic Principle and Architecture](#)

Nano Communication Networks

- [Nanonetworks](#)

Nano-electromagnetic Communications

- [Nanoscale Terahertz Communications](#)

Nanonetworks

Ian F. Akyildiz¹, Josep M. Jornet²,
and Massimiliano Pierobon³

¹Broadband Wireless Networking (BWN)
Laboratory, Georgia Institute of Technology,
Atlanta, GA, USA

²University at Buffalo, The State University of
New York, Buffalo, NY, USA

³University of Nebraska-Lincoln, Lincoln, NE,
USA

Synonyms

[Nano communication networks](#)

Definition

A nanonetwork is a set of interconnected nanomachines, i.e., devices that integrate nanometer-scale components, with computing, data storing, sensing, and actuation capabilities. Nanonetworks are expected to expand the capabilities of single nanomachines both in terms of complexity and range of operation by allowing them to coordinate, share, and fuse information. Nanonetworks enable new applications of nanotechnology in the biomedical field, environmental research, military technology, and industrial and consumer goods applications.

Historical Background

Nanotechnology and biotechnology are providing the engineering community with a new set of tools to control matter and program functions at the atomic and molecular scale. At this scale, novel nanomaterials, structures, and systems show new properties that cannot be realized at the microscopic level. By leveraging such properties, devices with unprecedented applications can be developed. Among others, nanosensors and nanoactuators, able to both monitor and control

physical, chemical, and biological processes at the nanoscale, have been realized (Wu and Qu 2015).

The realization of nanomachines with computation, data storage, and communication, in addition to sensing and actuation capabilities, is enabling transformative applications of nanotechnology in diverse fields (Akyildiz et al. 2008). By means of communication, nanomachines will be able to complete more complex tasks in a distributed manner. In addition, the integration of resulting nanonetworks with macronetworks and eventually the Internet leads to the ultimate cyberphysical systems, known as the Internet of Nano Things (Akyildiz and Jornet 2010) and the Internet of Bio-Nano Things (Akyildiz et al. 2015).

There are three paths toward developing nanomachines and enabling the communication between them, namely, *nanomaterial-based nanoscale electromagnetic networks*, *biomolecular communication networks*, and *hybrid nanonetworks*. Next, the state of the art is reviewed, and open challenges are identified for each approach.

Nanomaterial-Based Nanoscale Electromagnetic Networks

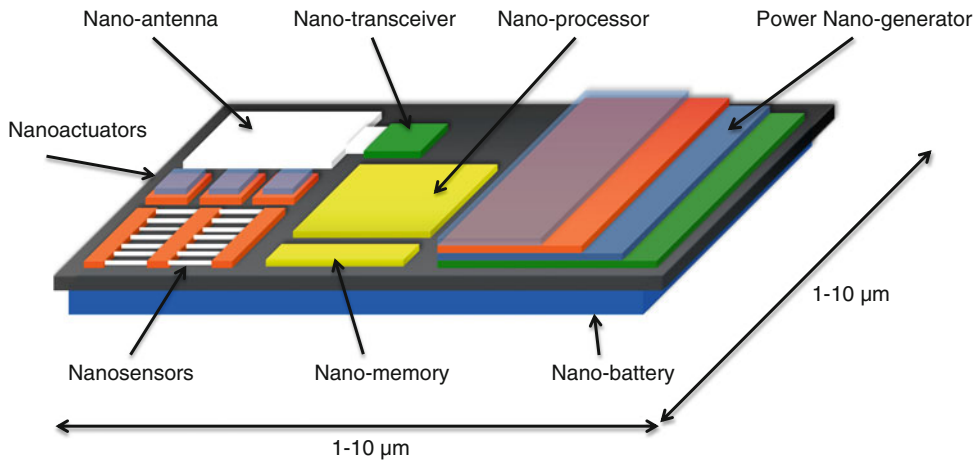
From Nanomaterials to Nanomachines

Nanomaterials are materials with any external dimension in the nanoscale (sizes range from approximately 1–100 nm) or having internal structure or surface structure in the nanoscale. Among others, graphene (Ferrari et al. 2015), a one-atom-thick planar sheet of bonded carbon atoms in a honeycomb crystal lattice, has been and still is one of the most popular nanomaterials. Graphene exhibits very unique mechanical, electrical, and optical properties. Among others, it is the world's thinnest and lightest material while being harder than diamond; it is a very good conductor with very high electron mobility; and, despite being one atom thick, it can absorb light efficiently. Graphene is the first of a kind, but not the only two-dimensional material. As

of today, more than 100 2D nanomaterials have reported, including molybdenum disulfide (MoS_2) (Splendiani et al. 2010) or hexagonal boron nitride (hBN) (Giovannetti et al. 2007), with different electronic properties. Moreover, the combination of different nanomaterials can lead to new nanostructures with novel properties (Geim and Grigorieva 2013).

By utilizing nanomaterials and the tools provided by nanotechnology, the different components which integrate a nanomachine (Fig. 1) can be realized, specifically,

- **Nano-processor** This is enabled by the development of smaller transistors. While 22 and 14 nm channel length transistor technologies are already commercial, the smallest transistor to date is based on a thin graphene strip made of just 10 by 1 carbon atoms, i.e., less than 1 nm long (Ponomarenko et al. 2008).
- **Nano-memory** Single-atom memories are at the basis of ultracompact data storing in nanomachines. For the time being, different single-atom memories based on different physical principles have been demonstrated, including electronic memories (Pla et al. 2012) and optical memories (Specht et al. 2011).
- **Power Nano-generator** In addition to nano-batteries (Brazier et al. 2008), energy-harvesting nano-systems or power nano-generators are needed for nanomachines to be able to continuously operate. Different energy-harvesting systems have been demonstrated to date, including mechanical energy harvesting through devices (Wang et al. 2017) or EM energy harvesting (Gadalla et al. 2014).
- **Nanosensor and Nanoactuator** Physical, chemical, and biological nanosensors and nanoactuators have been developed by using graphene and other nanomaterials. These are able to detect and measure magnitudes at the nanoscale, including the physical characteristics of structures just a few nanometers in size or the presence of biological agents such as virus or cancerous cells (Wu and Qu 2015).



Nanonetworks, Fig. 1 Conceptual architecture of a nanomachine

- Nano-transceiver and Nano-antenna**
 Nanomachines require miniature transceivers and antennas to communicate. On the one hand, noble metals can be utilized to fabricate plasmonic nano-antennas, which are just hundreds of nanometers in length and can efficiently operate at infrared and visible optical frequencies (Nafari and Jornet 2017). These require similarly small nano-lasers (Feng et al. 2014) and nano-photodetectors (Luo et al. 2016) to operate. On the other hand, graphene can be utilized to fabricate miniature nano-transceivers (Jornet and Akyildiz 2014b) and nano-antennas (Jornet and Akyildiz 2013), which can efficiently operate in the terahertz (THz) band (0.1–10 THz), i.e., at a frequency much lower than their metallic counterparts. These bring many opportunities for communication in nanonetworks, as it is explained next.

Besides the fabrication of each one of the components, there are several challenges related to their integration in a single device. While self-assembly techniques, including DNA scaffolding (Maune et al. 2010), are being developed, the utilization of a single nanomaterial or arrangement of nanomaterials to fabricate the

different components at once will drastically simplify the realization of nanomachines.

Terahertz and Optical Nanoscale Communication

The very small size of nano-antennas imposes the use of very high frequencies for EM wireless communication among nanomachines, ranging from the THz band to the infrared and visible optical frequency bands. This introduces multiple challenges and opportunities for communication and networking:

- Channel Modeling** There are three main phenomena which affect the propagation of EM waves at THz band and optical frequencies, namely, spreading, absorption, and scattering. Common to any wireless communication system, the *spreading loss* refers to attenuation due to the expansion of the wave front as it propagates through the medium. The *scattering loss* refers to the signal loss due to redirection (reflection, diffusion) of EM waves as they propagate through a medium with obstacles, which depends on the obstacle size in terms of wavelength. At THz frequencies, the main problem is posed by molecular absorption (Jornet and Akyildiz 2011; Han et al. 2015), whereas at optical frequencies, scattering is the main challenge

to overcome (Johari and Jornet 2018). In both cases, the channel can support beyond multi-GHz transmission bandwidths.

- **Physical Layer** The capabilities of nanomachines and the peculiarities of the channel need to be taken into account when designing the physical layer. In terms of *modulation*, the transmission of 100-femtosecond-long pulses spread in time has been proposed as a way to enable ultra-broadband communication in nanonetworks (Jornet and Akyildiz 2014a). These pulses have their main frequency components in the THz band (between 0.5 and 4 THz), can be generated with nano-transceivers, and are already widely used in THz sensing applications. Similarly, new *channel coding* strategies are needed to overcome channel errors resulting from weak signals, noise, and multiuser interference, such as low-weight channel codes (Jornet 2014). Independently of the modulation and coding strategy, physical-layer synchronization is a major bottleneck to overcome, due to the very low power and very short duration of the symbols.
- **Link Layer and Above** The very large available bandwidth at THz and optical frequencies reduces the need for the nodes to contend for the channel. However, the energy fluctuations, due to the use of energy-harvesting systems, require the development of link-layer synchronization and *medium access control (MAC)* algorithms for both ad hoc and infrastructure nanonetworks (Wang et al. 2013). Moreover, the very large number of nanomachines in nanonetworks requires the development of new *addressing* schemes, beyond IPv6, which leverage the hierarchical nature of nanonetworks and built upon compressed-addressing concepts. In addition, due to the limited transmission range of nanomachines, *multi-hop relaying and routing* strategies need to be developed (Pierobon et al. 2014; Tsioliariidou et al. 2015). Ultimately, ensuring *end-to-end reliability* will require the revision of well-established concepts at the transport layer.

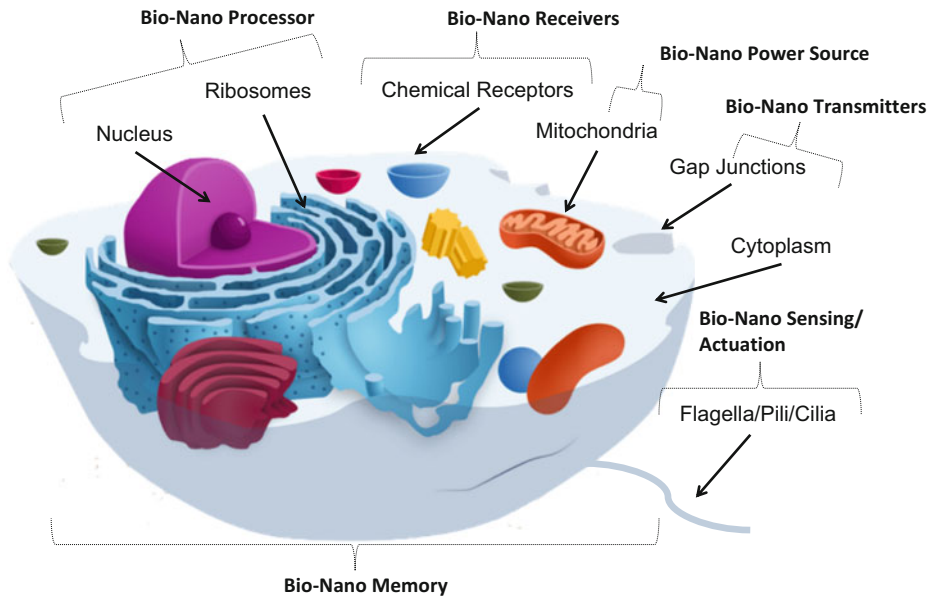
Biomolecular Communication Networks

Bio-nanomachines

Bio-nanomachines are uniquely identifiable basic structural and functional units that operate and interact within the biological environment. Stemming from biological cells, the basic units of life, and enabled by synthetic biology and nanotechnology, they are based on the control of biological processes (Akyildiz et al. 2015). Biotechnology, through DNA sequencing and synthesis tools, is enabling the reading and writing of cells' genetic code (Kahl and Endy 2013), providing open access to the set of structural and functional instructions at the basis of life. At the same time, nanotechnology is expanding these directions toward the assembly of artificial cells (Wu and Tan 2014).

The different components of a bio-nanomachine can be abstracted from the main components of a biological cell (Fig. 2) as follows:

- **Bio-Nano Processor** This corresponds to the molecular machinery that, from the DNA molecules (contained in the nucleus) to the so-called transcription and translation (at the ribosomes), generates protein molecules with instruction-dependent types and concentrations.
- **Bio-Nano Memory** This is an abstraction of the chemical content of the cytoplasm, i.e., the interior of the cell, comprised of molecules synthesized by the cell as a result of DNA instructions, and other molecules or structures, e.g., vesicles, exchanged with the external environment, and organelles, including the nucleus in case of eukaryotic cell, as shown in Fig. 2.
- **Bio-Nano Power Source** This is the reservoir of adenosine triphosphate (ATP) molecule, which is synthesized by the cell from energy supplied from the external environment in various forms, and provides the chemical energy necessary for the cell biochemical reactions to



Nanonetworks, Fig. 2 Main components of a eukaryotic cell (with nucleus) and their abstraction as components of a bio-nanomachine

take place. In the case of eukaryotic cells, ATP production occurs in mitochondria.

- **Bio-Nano Sensing and Actuation** This corresponds to the capability of a cell to chemically recognize external molecules or physical stimuli, e.g., light or mechanical stress, and to change the chemical characteristics of the environment or mechanically interact through moving elements, such as flagella, pili, or cilia.
- **Bio-Nano Transceivers** These are abstractions of specific chains of chemical reactions, i.e., signaling pathways, through which cells exchange information with each other and the environment.

Biomolecular Communications

Individual biological cells and cell populations perceive, elaborate, and exchange information through the propagation of molecules and their chemical reactions (Payne and You 2014), which can be abstracted according to the paradigm of molecular communication

(MC) (Akyildiz et al. 2011). The heterogeneity of biological MC mechanisms and the fundamental differences with respect to classical communication scenarios pose the following unique challenges and opportunities for the realization of bio-nanomachine interconnections into MC networks:

- **Channel Modeling** Several different MC mechanisms can be observed in nature that do not behave as the channels previously considered when engineering communication systems. In particular, the peculiarities of channels based on the simplest mechanism of molecular diffusion in terms of path loss and signal distortion (Pierobon and Akyildiz 2010), and their associated noise sources (Pierobon and Akyildiz 2011; Noel et al. 2014), necessitated completely novel interference (Pierobon and Akyildiz 2011), channel estimation (Jamali et al. 2016), and channel capacity (Pierobon and Akyildiz 2013; Atakan and Akan 2010) analyses. Other simple MC mechanisms have been

considered, such as molecular motors (Moore et al. 2006), gap junctions (Nakano et al. 2007; Kilinc and Akan 2012; Barros et al. 2015), and chemotaxis (Gregori and Akyildiz 2010), along with more complex MC scenarios, such as the cardiovascular system (Chahibi et al. 2013), while more fundamental communication theoretical aspects related to molecular information have been treated in Akan et al. (2017).

- **Physical Layer** The design of physical links between bio-nanomachines has to take into account different options to realize information encoding and modulation in MC systems, along with their theoretical performance (Lu et al. 2015; Murin et al. 2017), and molecular mechanisms to optimize signal detection and decoding (Kilinc and Akan 2013; Mahfuz et al. 2014, 2015; Chou 2015; Li et al. 2016; Jamali et al. 2018). Ultimately, the peculiarities of the underlying biological cells and the tools for their engineering should be accounted for. In this direction, synthetic-biology-enabled analog communications have been studied in Pierobon (2014) while digital-like communications in Unluturk et al. (2015). Channel coding strategies adapted to the efficiency of analog processing in cells have been suggested in Marcone et al. (2017).
- **Link Layer and Above** The definition of network architectures and protocols on top of the aforementioned biological MC systems will be particularly challenging given the nonlinear nature of many biochemical phenomena (Mian and Rose 2011) and the shared nature of the media, such as fluids (Deng et al. 2017). Optimization in terms of energy consumption for information propagation within these networks will be also an important aspect (Qiu et al. 2017), since cells have limited resources. A further challenge will be the interconnection of heterogeneous networks, i.e., composed of different types of bio-nanomachines and/or based on different biological MC mechanisms. The realization of interfaces between the electrical domain and the

biochemical domain will be the ultimate frontier to create a seamless interconnection between the biological environment and the Internet.

Hybrid Nanonetworks

Beyond nanomaterial-based nanomachines or synthetically engineered living cells, hybrid nano-bio-engineered nanomachines and nanonetworks will exhibit unique capabilities and enable applications, even yet to be envisioned. Among others, *optogenetics*, i.e., the use of light to control and monitor genetically engineered living cells, is at the basis of fundamental studies in developmental and neurodegenerative diseases as well as of brain machine interfaces, which can restore and even enhance functional abilities and cognition (Wirdatmadja et al. 2017). Beyond optogenetics, *optogenomics*, i.e., the control of gene expression with light, can lead to unprecedented control of biological processes. For all these, hybrid nanomachines, able to manipulate and interact with electromagnetic (THz, optical) and molecular signals, need to be developed.

Key Applications

The potential applications of the nanonetworks can be classified in four main areas: *biomedical applications*, including intrabody health monitoring and drug delivery systems, immune system support mechanisms, and artificial biohybrid implants; *industrial and consumer goods applications*, development of intelligent functionalized materials and fabrics, new manufacturing processes and distributed quality control procedures, and food and water quality control systems; *environmental applications*, biological and chemical nanosensor networks for pollution control, biodegradation assistance, and animal and biodiversity control; and *military applications*, nuclear, biological, and chemical defenses and nano-functionalized equipment and human enhancement.

Cross-References

- [Brain-Machine Interfaces](#)
- [Nanoscale Terahertz Communications](#)

References

- Akan OB, Ramezani H, Khan T, Abbasi NA, Kuscü M (2017) Fundamentals of molecular information and communication science. *Proc IEEE* 105(2):306–318
- Akyildiz IF, Jornet JM (2010) The Internet of nano-things. *IEEE Wirel Commun Mag* 17(6):58–63
- Akyildiz IF, Brunetti F, Blazquez C (2008) Nanonetworks: a new communication paradigm. *Comput Netw (Elsevier)* 52(12):2260–2279
- Akyildiz IF, Jornet JM, Pierobon M (2011) Nanonetworks: a new frontier in communications. *Commun ACM* 54(11):84–89
- Akyildiz IF, Pierobon M, Balasubramaniam S, Koucheryavy Y (2015) The internet of bio-nano things. *IEEE Commun Mag* 53(3):32–40
- Atakan B, Akan OB (2010) Deterministic capacity of information flow in molecular nanonetworks. *Nano Commun Netw* 1(1):31–42
- Barros MT, Jennings B, Balasubramaniam S (2015) Comparative end-to-end analysis of Ca^{2+} -signaling-based molecular communication in biological tissues. *IEEE Trans Commun* 63(12):5128–5142
- Brazier A, Dupont L, Dantras-Laffont L, Kuwata N, Kawamura J, Tarascon JM (2008) First cross-section observation of an all solid-state lithium-ion “nanobattery” by transmission electron microscopy. *Chem Mater* 20(6):2352–2359
- Chahibi Y, Pierobon M, Song SO, Akyildiz I (2013) A molecular communication system model for particulate drug delivery systems. *IEEE Trans Biomed Eng* 60(12):3468–3483
- Chou CT (2015) Maximum a-posteriori decoding for diffusion-based molecular communication using analog filters. *IEEE Trans Nanotechnol* 14(6):1054–1067
- Deng Y, Noel A, Guo W, Nallanathan A, Elkashlan M (2017) Analyzing large-scale multiuser molecular communication via 3D stochastic geometry. *IEEE Trans Mol Biol Multiscale Commun* 3(2):118–133
- Feng L, Wong ZJ, Ma RM, Wang Y, Zhang X (2014) Single-mode laser by parity-time symmetry breaking. *Science* 346(6212):972–975
- Ferrari AC, Bonaccorso F, Fal’Ko V, Novoselov KS, Roche S, Bøggild P, Borini S, Koppens FH, Palermo V, Pugno N et al (2015) Science and technology roadmap for graphene, related two-dimensional crystals, and hybrid systems. *Nanoscale* 7(11):4598–4810
- Gadalla M, Abdel-Rahman M, Shamim A (2014) Design, optimization and fabrication of a 28.3 Thz nano-rectenna for infrared detection and rectification. *Sci Rep* 4:4270
- Geim AK, Grigorieva IV (2013) Van der waals heterostructures. *Nature* 499(7459):419
- Giovannetti G, Khomyakov PA, Brocks G, Kelly PJ, Van Den Brink J (2007) Substrate-induced band gap in graphene on hexagonal boron nitride: ab initio density functional calculations. *Phys Rev B* 76(7):073103
- Gregori M, Akyildiz IF (2010) A new nanonetwork architecture using flagellated bacteria and catalytic nanomotors. *IEEE J Sel Areas Commun (JSAC)* 28(4):612–619
- Han C, Bicen AO, Akyildiz I (2015) Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band. *IEEE Trans Wirel Commun* 14(5):2402–2412
- Jamali V, Ahmadzadeh A, Jardin C, Sticht H, Schober R (2016) Channel estimation for diffusive molecular communications. *IEEE Trans Commun* 64(10):4238–4252
- Jamali V, Farsad N, Schober R, Goldsmith A (2018) Non-coherent detection for diffusive molecular communication systems. *IEEE Trans Commun* 66(6):2515–s2531
- Johari P, Jornet JM (2018) Nanoscale optical wireless channel model for intra-body communications: geometrical, time, and frequency domain analyses. *IEEE Trans Commun* 66(4):1579–1593
- Jornet JM (2014) Low-weight error-prevention codes for electromagnetic nanonetworks in the terahertz band. *Nano Commun Netw (Elsevier)* 5(1–2):35–44
- Jornet JM, Akyildiz IF (2011) Channel modeling and capacity analysis of electromagnetic wireless nanonetworks in the terahertz band. *IEEE Trans Wirel Commun* 10(10):3211–3221
- Jornet JM, Akyildiz IF (2013) Graphene-based plasmonic nano-antenna for terahertz band communication in nanonetworks. *IEEE JSAC Special Issue Emerg Technol Commun* 12(12):685–694
- Jornet JM, Akyildiz IF (2014a) Femtosecond-long pulse-based modulation for terahertz band communication in nanonetworks. *IEEE Trans Commun* 62(5):1742–s1754
- Jornet JM, Akyildiz IF (2014b) Graphene-based plasmonic nano-transceiver for terahertz band communication. In: *Proceeding of European conference on antennas and propagation (EuCAP)*, The Hague
- Kahl LJ, Endy D (2013) A survey of enabling technologies in synthetic biology. *J Biol Eng* 7(1):13
- Kilinc D, Akan OB (2012) An information theoretical analysis of nanoscale molecular gap junction communication channel between cardiomyocytes. *IEEE Trans Nanotechnol* 12(2):129–136
- Kilinc D, Akan OB (2013) Receiver design for molecular communication. *IEEE J Sel Areas Commun* 31(12):705–714
- Li B, Sun M, Wang S, Guo W, Zhao C (2016) Local convexity inspired low-complexity non-coherent signal detector for molecular communications. *IEEE Trans Commun* 64(5):2079–2091

- Lu Y, Higgins MD, Leeson MS (2015) Comparison of channel coding schemes for molecular communications systems. *IEEE Trans Commun* 63(11):3991–4001
- Luo LB, Zou YF, Ge CW, Zheng K, Wang DD, Lu R, Zhang TF, Yu YQ, Guo ZY (2016) A surface plasmon enhanced near-infrared nanophotodetector. *Adv Opt Mater* 4(5):763–771
- Mahfuz MU, Makrakis D, Mouftah HT (2014) A comprehensive study of sampling-based optimum signal detection in concentration-encoded molecular communication. *IEEE Trans Nanobiosci* 13(3):208–222
- Mahfuz MU, Makrakis D, Mouftah HT (2015) A comprehensive analysis of strength-based optimum signal detection in concentration-encoded molecular communication with spike transmission. *IEEE Trans Nanobiosci* 14(1):67–83
- Marccone A, Pierobon M, Magarini M (2017) A parity check analog decoder for molecular communication based on biological circuits. In: *In Proceedings of the IEEE international conference on computer communications (INFOCOM)*
- Maune HT, Han Sp, Barish RD, Bockrath M, Goddard WA III, Rothmund PW, Winfree E (2010) Self-assembly of carbon nanotubes into two-dimensional geometries using dna origami templates. *Nature Nanotechnol* 5(1):61
- Mian IS, Rose C (2011) Communication theory and multicellular biology. *Integr Biol* 3(4):350–367
- Moore M, Enomoto A, Nakano T, Egashira R, Suda T, Kayasuga A, Kojima H, Sakakibara H, Oiwa K (2006) A design of a molecular communication system for nanomachines using molecular motors. In: *Fourth annual IEEE international conference on pervasive computing and communications workshops, Pisa*, pp 6–12
- Murin Y, Farsad N, Chowdhury M, Goldsmith A (2017) Time-slotted transmission over molecular timing channels. *Nano Commun Netw* 12:12–24
- Nafari M, Jornet JM (2017) Modeling and performance analysis of metallic plasmonic nano-antennas for wireless optical communication in nanonetworks. *IEEE Access* 5:6389–6398
- Nakano T, Suda T, Koujin T, Haraguchi T, Hiraoka Y (2007) Molecular communication through gap junction channels: system design, experiments and modeling. In: *Proceedings of 2nd international conference on bio-inspired models of network, information and computing systems (Bionetics)*, Budapest, pp 139–146
- Noel A, Cheung KC, Schober R (2014) A unifying model for external noise sources and ISI in diffusive molecular communication. *IEEE J Sel Areas in Commun* 32(12):2330–2343
- Payne S, You L (2014) Engineered cell-cell communication and its applications. *Adv Biochem Eng Biotechnol* 146:97–121
- Pierobon M (2014) A systems-theoretic model of a biological circuit for molecular communication in nanonetworks. *Nano Commun Netw (Elsevier)* 5(1–2):25–34
- Pierobon M, Akyildiz IF (2010) A physical end-to-end model for molecular communication in nanonetworks. *IEEE J Sel Areas Commun (JSAC)* 28(4):602–611
- Pierobon M, Akyildiz IF (2011) Diffusion-based noise analysis for molecular communication in nanonetworks. *IEEE Trans Signal Process* 59(6):2532–2547
- Pierobon M, Akyildiz IF (2013) Capacity of a diffusion-based molecular communication system with channel memory and molecular noise. *IEEE Trans Inf Theory* 59(2):942–954
- Pierobon, Massimiliano, et al. (2014) A routing framework for energy harvesting wireless nanosensor networks in the Terahertz Band. *Wireless networks* 20(5): 1169–1183. <https://link.springer.com/article/10.1007/s11276-013-0665-y>
- Pla JJ, Tan KY, Dehollain JP, Lim WH, Morton JJ, Jamieson DN, Dzurak AS, Morello A (2012) A single-atom electron spin qubit in silicon. *Nature* 489(7417):541
- Ponomarenko LA, Schedin F, Katsnelson MI, Yang R, Hill EW, Novoselov KS, Geim AK (2008) Chaotic Dirac billiard in graphene quantum dots. *Science* 320(5874):356–358
- Qiu S, Haselmayr W, Li B, Zhao C, Guo W (2017) Bacterial relay for energy efficient molecular communications. *IEEE Trans NanoBiosci* 7(16):555–562
- Specht HP, Nölleke C, Reiserer A, Uphoff M, Figueroa E, Ritter S, Rempe G (2011) A single-atom quantum memory. *Nature* 473(7346):190
- Splendiani A, Sun L, Zhang Y, Li T, Kim J, Chim CY, Galli G, Wang F (2010) Emerging photoluminescence in monolayer MOS₂. *Nano Lett* 10(4):1271–1275
- Tsioliariidou A, Liaskos C, Ioannidis S, Pitsillides A (2015) Corona: a coordinate and routing system for nanonetworks. In: *Proceedings of the second annual international conference on nanoscale computing and communication*. ACM, New York, p 18
- Unluturk B, Bicen A, Akyildiz I (2015) Genetically engineered bacteria-based biotransceivers for molecular communication. *IEEE Trans Commun* 63(4):1271–1281
- Wang P, Jornet JM, Abbas Malik M, Akkari N, Akyildiz IF (2013) Energy and spectrum-aware MAC protocol for perpetual wireless nanosensor networks in the terahertz band. *Ad Hoc Netw (Elsevier)* 11(8):2541–2555
- Wang ZL, Jiang T, Xu L (2017) Toward the blue energy dream by triboelectric nanogenerator networks. *Nano Energy* 39:9–23
- Wirdatmadja SA, Barros MT, Koucheryavy Y, Jornet JM, Balasubramaniam S (2017) Wireless optogenetic nanonetworks for brain stimulation: device model and charging protocols. *IEEE Trans NanoBiosci* 16(8):859–872
- Wu F, Tan C (2014) The engineering of artificial cellular nanosystems using synthetic biology approaches. *WIREs Nanomed Nanobiotechnol* 6(4):369–383
- Wu L, Qu X (2015) Cancer biomarker detection: recent achievements and challenges. *Chem Soc Rev* 44(10):2963–2997

Nano-networks

► Applications of Molecular Communication Systems

Nanoscale Terahertz Communications

Chong Han¹, Josep Miquel Jornet², and Ian F. Akyildiz³

¹University of Michigan-Shanghai Jiao Tong University Joint Institute (UM-SJTU JI), Shanghai Jiao Tong University, Shanghai, China

²Ultra-Broadband Nano Communication and Networking Laboratory, University at Buffalo, The State University of New York, Buffalo, NY, USA

³Broadband Wireless Networking (BWN) Laboratory, Georgia Institute of Technology, Atlanta, GA, USA

Synonyms

Electromagnetic nanonetworks; Nano-electromagnetic Communications; Terahertz-band nano-communications

Definition

Nanoscale terahertz communications refer to the wireless technology that enables the exchange of information and coordination between nanomachines in the electromagnetic frequency range spanning from 100 GHz up to 10 THz.

Historical Background

Nanotechnology is providing the engineering community with a new set of tools to create miniature machines, just a few cubic micrometers in size, with sensing, actuation, computation, and data storing capabilities (Ferrari et al. 2015). By

means of communication, such nanomachines will be able to accomplish more complex tasks in a distributed manner. Nanonetworks (Akyildiz et al. 2011), i.e., networks of nanomachines, enable transformative applications in the biomedical, environmental, security, defense, and consumer field. While different communication technologies are being considered to enable nanonetworks, in this entry, the focus is on the use of electromagnetic (EM) signals at THz-band frequencies.

Why Communication in the Terahertz Band?

Enabling nanoscale EM communication requires the development of nanoscale transceivers and antennas. The miniaturization of a metallic antenna would impose the use of very high resonant frequencies. For example, a 1-micrometer-long dipole antenna would expectedly resonate at 150 THz, i.e., in the near infrared. However, at infrared and visible optical frequencies, metals do not behave as perfect electric conductors (PEC) anymore, and thus, common assumptions in antenna theory need to be revised (Dorfmueller et al. 2010; Nafari and Jornet 2017). Moreover, in addition to the antenna, optical nano-transmitters (e.g., nano-lasers) and optical nano-receivers (e.g., nanophotodetectors) will be needed to enable nanoscale optical communication. Even when those become available, the very high propagation loss at optical frequencies combined with the limited power of nanomachines can compromise the feasibility of nanonetworks.

To overcome these limitations, graphene can be utilized to develop novel nano-transceivers and nano-antennas able to operate at lower frequencies than their metallic counterparts. Graphene is a two-dimensional carbon crystal that has excellent electrical conductivity, making it very well suited for propagating extremely high-frequency electrical signals (Ferrari et al. 2015). Moreover, graphene supports the propagation of THz surface plasmon polariton (SPP) waves at room temperature (Koppens et al. 2011; Vakil and Engheta 2011). SPP waves are surface-confined EM waves that appear at the interface of a metal (graphene

in this case) and a dielectric (air or any other substrate on which graphene is deposited). The propagation properties of SPP waves, which can be electronically tuned, determine the behavior of graphene-based plasmonic nano-devices.

In Jornet and Akyildiz (2010, 2013), the use of graphene to build plasmonic nano-antennas was proposed for the first time. The fundamental idea is to leverage the reduced propagation speed of SPP waves on graphene to design resonant antennas which are physically small but equivalently electrically large. As a result, a 1-micrometer-long nano-antenna, just tens to hundreds of nanometers wide, can efficiently radiate at frequencies in the THz band, i.e., at a frequency two orders of magnitude lower than metallic antennas with the same size. In successive works, it has been shown that the resonant frequency of the antennas can be electrically tuned (Tamagnone et al. 2012), and different shapes and designs have been considered (Dragoman et al. 2015).

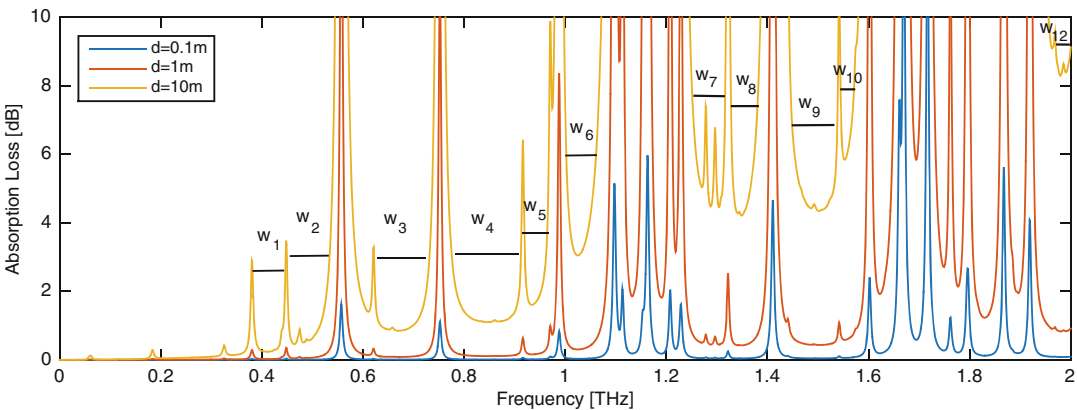
In addition to the nano-antenna, mechanisms to generate, modulate, detect, and demodulate the THz signals are needed. Different mechanisms have been considered to date. In Jornet and Akyildiz (2014b), a THz plasmonic signal source and detector based on a hybrid graphene III–V semiconductor HEMT was proposed. Compared to existing plasmonic sources and detectors, the generated plasma wave is not directly radiated but utilized to launch an SPP wave on the graphene layer that extends toward the nano-antenna. Other options include the use of photoconductive antennas, which downconvert optical radiation into THz-band radiation (Yardimci et al. 2015), plasmonic grating structures (Otsuji et al. 2013), or quantum cascade lasers (Lu et al. 2016). For all these, the THz band is envisioned as the frequency band of communication of nanomachines.

Terahertz-Band Channel Modeling

Characterizing the THz channel is a necessary step toward enabling nanoscale THz communications. While the frequency regions immediately below (i.e., microwave and millimeter-wave) and above (i.e., infrared and visible optical) this band have been extensively investigated, this

is one of the least-explored frequency bands in the EM spectrum. In Jornet and Akyildiz (2011), the first channel model for the entire THz band was developed, by utilizing tools from radiative transfer theory, electromagnetics and communication theory, and leveraging the contents of the high-resolution transmission molecular absorption (HITRAN) database. There are two main phenomena affecting the propagation of EM signals at THz frequencies in free space, namely, spreading and molecular absorption. The *spreading loss*, which is common to any wireless communication system, accounts for the attenuation due to the expansion of the wave as it propagates through the medium, and it depends only on the signal frequency and the transmission distance. The *absorption loss* accounts for the attenuation that a propagating wave suffers because of molecular absorption, i.e., the process by which part of the wave energy is converted into internal kinetic energy in some of the molecules that are found in the medium. This depends on the concentration and the particular mixture of molecules encountered along the path. Different types of molecules have different resonant frequencies, and in addition, the absorption at each resonance is not confined to a single frequency but spread over a range of frequencies. As a result, the terahertz channel is very frequency selective. A complete list of absorption-defined spectral windows is provided in Fig. 1, with carrier frequency f_c , 3-dB bandwidth B_{3dB} , molecular absorption loss A_{abs} , and total path loss A , over the distances at 0.1, 1, and 10 m. While, for long distances, the effort is currently on characterizing individual transmission windows (e.g., 300 GHz Priebe et al. 2011), it is clear that for nanoscale THz communications, the entire THz band can be considered as a single transmission window.

The presence of obstacles further complicates the propagation of THz waves. Any object whose size is beyond a few wavelengths (i.e., a few millimeters or more) behaves as an obstacle for THz signals, which can be reflected, scattered, or diffracted, leading to multipath propagation. In this direction, a unified multi-ray channel



Name	f_c [THz]	B_{3dB} [GHz]	A_{abs} [dB]	A [dB]	Name	f_c [THz]	B_{3dB} [GHz]	A_{abs} [dB]	A [dB]
w_1	0.41	65.61	0.24	105.02	w_7	1.3	38.15	4.46	121.39
w_2	0.49	86.21	0.46	106.86	w_8	1.35	51.12	4.21	119.41
w_3	0.66	152.59	0.85	109.76	w_9	1.49	91.55	4.34	120.47
w_4	0.84	141.91	1.10	112.07	w_{10}	1.56	28.99	5.84	122.37
w_5	0.94	47.3	1.61	113.62	w_{11}	1.83	25.18	12.31	130.03
w_6	1.03	57.98	3.05	115.85	w_{12}	1.98	56.46	8.30	126.69

Nanoscale Terahertz Communications, Fig. 1 Molecular absorption-defined spectral windows in the THz band, specifically for distance over 10 m. Windows

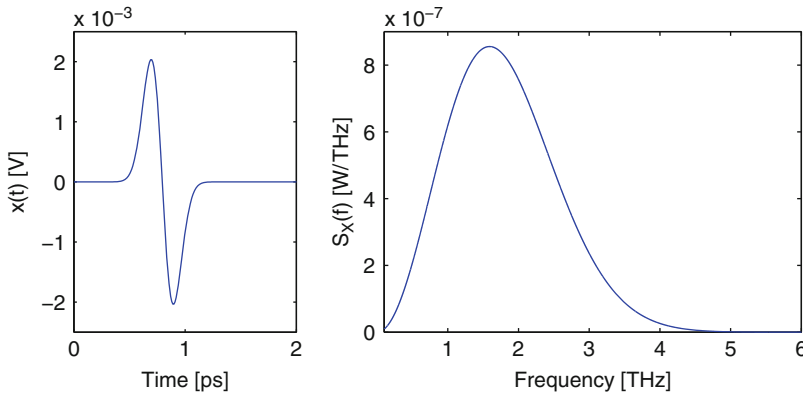
are shown in the top figure (w_{11} is missing due to its high absorption loss), while complete windows are listed in the bottom table

model for the THz band was first provided in Han et al. (2015), based on ray tracing techniques and accounting the propagation models for the line of sight (LoS), reflected, scattered, and diffracted paths. The developed theoretical model was validated with the experimental measurements at 60 and 300 GHz from the literature. Then, using the developed propagation models, an in-depth analysis on the THz channel characteristics is carried out, in terms of the distance-varying and frequency-selective spectral windows, coherence bandwidth, wideband channel capacity, and the temporal broadening effects. The principles to develop efficient multi-ray model are summarized here. When LoS is available, the direct ray dominates the received signal energy, while the reflected rays play a dominant role when LoS is absent. As the operating frequency increases, the surface is seen to be rougher, and hence, more power is scattered out of the specular direction. Hence, scattered rays are very important and have to be included in the ray tracing model for both LoS and NLoS conditions. Furthermore, the diffraction path can be ignored in general, only except when the receiver is in

the very closed region near the incident shadow boundary. Further stochastic modeling studies are presented by Hossain et al. (2017) and He et al. (2017).

Communication Mechanisms

The very large bandwidth supported by the THz-band channel over short distances enables new modulation strategies, which are also well suited for the limited capabilities and energy constraints of nanomachines. For example, ultra-low-energy sub-picosecond-long pulses can be utilized. These pulses, whose major frequency components are between 0.5 and 4 THz (see Fig. 2), are already being used in several applications such as THz spectroscopy and imaging systems. In this direction, TS-OOK, a new communication scheme based on the transmission of 100-femtosecond-long Gaussian pulses by following an on-off keying modulation spread in time, was first proposed in Jornet and Akyildiz (2014a). In this scheme, a logical “1” is transmitted by using a 100-femtosecond-long pulse, and a logical “0” is transmitted as silence, i.e., the device remains silent when a logical



Nanoscale Terahertz Communications, Fig. 2 Time and frequency responses of a femtosecond-long pulse for THz-band communications

zero is transmitted. To benefit from this transient behavior of molecular absorption and account for the temporal broadening effect, the time between symbols needs to be much longer than the pulse duration (e.g., a few tens of picoseconds), and, for convenience, it is a fixed parameter in the proposed scheme. Note that under TS-OOK, many nano-devices can simultaneously transmit without creating interference. Indeed, due to the fact that the time between transmissions is expectedly much longer than the pulse duration, several nanomachines can concurrently use the channel without affecting each other. To fully exploit this phenomenon, low-weight channel coding strategies have been proposed (Jornet 2014).

Furthermore, as an extension from the carrierless to the carrier-based modulation scheme, a multi-wideband waveform design for distance-adaptive THz-band communications is developed by Han et al. (2016), which includes the features of the pseudorandom time-hopping sequence and the polarity randomization. To cope with the unique characteristics and improve the distance, the pulse waveform design in the distance-adaptive multi-wideband system for the THz band includes the development of the waveform model and the dynamical adaptation of the rate and the transmit power on each sub-window. An optimization framework is formulated to solve these design parameters, with the aim is to maximize the communication distance while satisfying the rate and the transmit

power constraints. Accurate synchronization for these ultra-short pulse communications is critical, for which a low-sampling-rate (i.e., fraction of Nyquist rate) timing acquisition algorithm is proposed, by extending the theory of sampling signals with finite rate of innovation in the communication context and exploiting the properties of the annihilating filter (Maravic and Vetterli 2005; Han et al. 2017). Interference and coverage analysis based on stochastic geometry method are studied (Yao et al. 2017), taking into account the nanonetwork density, THz frequency band, and interference strength.

Network Protocols

The interconnection of nanomachines with existing wireless communication networks and ultimately the Internet (Akyildiz and Jornet 2010) requires the development of nano-micro interface devices, to bridge the nanoscale and micro-scale systems.

Networking protocols for nanoscale THz networks need to take into account the capabilities of nanomachines and the peculiarities of the THz channel. At the link layer, the very large bandwidth provided by the THz channel and the intrinsic multiuser interleaving ability of TS-OOK simplifies the regulation of the channel access. Nevertheless, energy limitations of nanomachines introduce many other challenges. In this direction, an energy and spectrum-aware medium access protocol for

nanoscale networks was proposed (Wang et al. 2013), to achieve fair, throughput, and lifetime optimal channel access by jointly optimizing the energy harvesting and consumption processes. In particular, the maximum allowable ratio between the transmission time and the energy harvesting time is derived, below which the nano-nodes can harvest more energy than the consumption, theoretically enabling perpetual data transmission. By utilizing the elasticity of the inter-symbol spacing of the pulse-based physical layer, multiple nano-nodes can transmit packets in parallel without causing transmission collisions, which forms a symbol-compression scheduling algorithm. In addition, a receiver-initiated MAC protocol for nanonetworks was presented in Mohrehkesh and Weigle (2014). Still at a very early stage, routing in nanonetworks requires further attentions.

Key Applications

- **In vivo health monitoring and drug delivery systems** Nanosensors can be implanted to operate inside the human body in real time and provide fast and accurate disease diagnosis. Furthermore, autonomous nanorobots can be used to serve drug delivery and even minimally invasive surgery (Elayan et al. 2017).
- **Nuclear, biological, and chemical defenses** Chemical and biological nanosensors can be used to detect harmful chemicals and biological weapons in a distributed manner, by precisely sensing the concentration as low as one molecule.
- **The Internet of Nano-Things** The interconnection of nanoscale machines with existing communication networks and ultimately Internet defines a truly cyber-physical system which known as the Internet of Nano-Things (IoNT) (Akyildiz and Jornet 2010), as one step forward from the Internet of Things.
- **Wireless network-on-chip (WiNoC) communications** The THz band can provide efficient and scalable means of inter-core communication in WiNoC networks, by using planar nano-antennas to create ultra-high-

speed links. More importantly, the use of graphene-based THz band communication (Abadal et al. 2013) would deliver inherent multicast and broadcast communication capabilities at the core level.

Cross-References

- [Channel Modeling of Microscale Terahertz Communication](#)
- [Nanonetworks](#)

References

- Abadal S, Alarcón E, Cabellos-Aparicio A, Lemme M, Nemirovsky M (2013) Graphene-enabled wireless communication for massive multicore architectures. *IEEE Commun Mag* 51(11):137–143
- Akyildiz IF, Jornet JM (2010) The Internet of nano-things. *IEEE Wirel Commun* 17(6):58–63
- Akyildiz IF, Jornet JM, Pierobon M (2011) Nanonetworks: a new frontier in communications. *Commun ACM* 54(11):84–89
- Dorfmueller J, Vogelgesang R, Khunsin W, Rockstuhl C, Etrich C, Kern K (2010) Plasmonic nanowire antennas: experiment, simulation, and theory. *Nano Lett* 10(9):3596–3603
- Dragoman M, Neculoiu D, Bunea AC, Deligeorgis G, Aldrigo M, Vasilache D, Dinescu A, Konstantinidis G, Mencarelli D, Pierantoni L et al (2015) A tunable microwave slot antenna based on graphene. *Appl Phys Lett* 106(15):153101
- Elayan H, Johari P, Shubair RM, Jornet JM (2017) Photothermal modeling and analysis of intra-body terahertz nanoscale communication. *IEEE Trans Nanobiosci* 16(8):755–763
- Ferrari AC, Bonaccorso F, Fal'Ko V, Novoselov KS, Roche S, Bøggild P, Borini S, Koppens FH, Palermo V, Pugno N et al (2015) Science and technology roadmap for graphene, related two-dimensional crystals, and hybrid systems. *Nanoscale* 7(11):4598–4810
- Han C, Bicen AO, Akyildiz IF (2015) Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band. *IEEE Trans Wirel Commun* 14(5):2402–2412
- Han C, Bicen AO, Akyildiz IF (2016) Multi-wideband waveform design for distance-adaptive wireless communications in the terahertz band. *IEEE Trans Signal Process* 64(4):910–922
- Han C, Akyildiz IF, Gerstaecker WH (2017) Timing acquisition and error analysis for pulse-based terahertz band wireless systems. *IEEE Trans Veh Technol* 66(11):10102–10113

- He D, Guan K, Fricke A, Ai B, He R, Zhong Z, Kasamatsu A, Hosako I, Kürner T (2017) Stochastic channel modeling for kiosk applications in the terahertz band. *IEEE Trans Terahertz Sci Technol* 7(5):502–513
- Hossain Z, Mollica C, Jornet JM (2017) Stochastic multipath channel modeling and power delay profile analysis for terahertz-band communication. In: *Proceedings of the 4th ACM international conference on nanoscale computing and communication*. ACM, New York, p 32
- Jornet JM (2014) Low-weight error-prevention codes for electromagnetic nanonetworks in the terahertz band. *Nano Commun Netw* (Elsevier) J 5(1–2):35–44
- Jornet JM, Akyildiz IF (2010) Graphene-based nano-antennas for electromagnetic nanocommunications in the terahertz band. In: *Proceedings of 4th European conference on antennas and propagation*, EUCAP, Barcelona, Spain
- Jornet JM, Akyildiz IF (2011) Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the terahertz band. *IEEE Trans Wirel Commun* 10(10):3211–3221
- Jornet JM, Akyildiz IF (2013) Graphene-based plasmonic nano-antenna for terahertz band communication in nanonetworks. *IEEE J Sel Areas Commun* 31(12):685–694
- Jornet JM, Akyildiz IF (2014a) Femtosecond-long pulse-based modulation for terahertz band communication in nanonetworks. *IEEE Trans Commun* 62(5):1742–1754
- Jornet JM, Akyildiz IF (2014b) Graphene-based plasmonic nano-transceiver for terahertz band communication. In: *Proceedings of European conference on antennas and propagation (EuCAP)*, The Hague, Netherlands
- Koppens FHL, Chang DE, Garcia de Abajo FJ (2011) Graphene plasmonics: a platform for strong light matter interactions. *Nano Lett* 11(8):3370–3377
- Lu Q, Wu D, Sengupta S, Slivken S, Razeghi M (2016) Room temperature continuous wave, monolithic tunable THz sources based on highly efficient mid-infrared quantum cascade lasers. *Sci Rep* 6:23595
- Maravic I, Vetterli M (2005) Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. *IEEE Trans Signal Process* 53(8):2788–2805
- Mohrehkesh S, Weigle MC (2014) Rih-MAC: receiver-initiated harvesting-aware MAC for nanonetworks. In: *Proceedings of the ACM international conference on nanoscale computing and communication*, Atlanta, USA
- Nafari M, Jornet JM (2017) Modeling and performance analysis of metallic plasmonic nano-antennas for wireless optical communication in nanonetworks. *IEEE Access* 5:6389–6398
- Otsuji T, Watanabe T, Boubanga Tombet S, Satou A, Knap W, Popov V, Ryzhii M, Ryzhii V (2013) Emission and detection of terahertz radiation using two-dimensional electrons in III–V semiconductors and graphene. *IEEE Trans Terahertz Sci Technol* 3(1):63–71
- Priebe S, Jastrow C, Jacob M, Kleine-Ostmann T, Schrader T, Kurner T (2011) Channel and propagation measurements at 300 ghz. *IEEE Trans Antennas Propag* 59(5):1688–1698
- Tamagnone M, Gomez-Diaz JS, Mosig JR, Perruisseau-Carrier J (2012) Reconfigurable terahertz plasmonic antenna concept using a graphene stack. *Appl Phys Lett* 101(21):214102
- Vakil A, Engheta N (2011) Transformation optics using graphene. *Science* 332(6035):1291–1294
- Wang P, Jornet JM, Malik MA, Akkari N, Akyildiz IF (2013) Energy and spectrum-aware MAC protocol for perpetual wireless nanosensor networks in the terahertz band. *Elsevier Ad Hoc Netw* 11(8):2541–2555
- Yao XW, Wang CC, Wang WL, Han C (2017) Stochastic geometry analysis of interference and coverage in terahertz networks. *Nano Commun Netw* 13:9–19
- Yardimci NT, Yang SH, Berry CW, Jarrahi M (2015) High-power terahertz generation using large-area plasmonic photoconductive emitters. *IEEE Trans Terahertz Sci Technol* 5(2):223–229

Narrowband Internet of Things

Chih-Wei Huang

Department of Communication Engineering,
National Central University, Taoyuan, Taiwan

Synonyms

NB-IoT

Definitions

The Narrowband Internet of Things (NB-IoT) is a major feature enhancement toward massive machine-type communications (mMTC) addressed in the Third Generation Partnership Project (3GPP) working group, which targets on devices requiring low data rate, low cost, and long battery life.

Historical Background

With the evolving toward a connected world with the Internet of Things (IoT), low-power wide-area (LPWA) network is developed to support devices that have low power consumption, low mobility, and long range. The LPWA category

includes both standardized and proprietary wireless technologies; NB-IoT is one of them standardized by the 3GPP. NB-IoT was first introduced by 3GPP Release 13 in June 2016. The performance enhancements in Release 13 (3GPP TR 45.820 [2017](#)) include (1) about 55,000 devices per cell under the suggested service model, (2) 164 + dB of coverage, (3) 10 or more years of battery life, and (4) latency in 10 s. In addition to legacy long-term evolution (LTE) service requests, NB-IoT defines control and user plane optimization for more efficient application data transmission (3GPP TS 23.401 [2017](#)). In 2017, the 3GPP Release 14 provides further enhancements in power efficiency, multicast support, and positioning.

Foundations

NB-IoT is designed to coexist with GSM and LTE systems in three operation modes. In the in-band mode, an LTE physical resource block (PRB) is used as a narrowband to transmit signaling and traffic data. In the guard-band mode, NB-IoT channels operate in the guard bands of LTE physical channels. In the standalone mode, the re-farmed GSM low-frequency bands, such as 700, 800, and 900 MHz, are used by NB-IoT. For the downlink transmission, NB-IoT adopts the orthogonal frequency-division multiple access (OFDMA) technology with 15 kHz subcarrier spacing like LTE. For the uplink transmission, the single-carrier frequency-division multiple access (SC-FDMA) technology is adopted supporting 15 and 3.75 kHz carrier spacing for single-tone and multitone transmissions, respectively. Frequency-division duplex technology with the half-duplex transmission is used to separate uplink and downlink frequency bands (3GPP TS 36.211 [2017](#)).

The NB-IoT radio resources are arranged into radio frames in the time domain and channels in the frequency domain. A radio frame consists of ten 1-ms-long subframes, and an NB-IoT channel contains 12 subcarriers reutilizing the LTE PRB structure. At the same time, physical channels

and signals are explicitly designed for narrow-band access (3GPP TS 36.213 [2017](#)).

For the downlink, three physical channels are defined to carry user data and control messages:

- Narrowband Physical Broadcast Channel (NPBCH) broadcasts the master information block (MIB) with system information. NPBCH occupies the first subframe of every radio frame consuming 10% of downlink resources.
- Narrowband Physical Downlink Control Channel (NPDCCH) carries UE-specific downlink control information (DCI) for data transmission, paging, and random access. The specification of NPDCCH is indicated by parameters $R_{\max}$ (the maximum DCI repetition number) and G (a system parameter).
- Narrowband Physical Downlink Shared Channel (NPDSCH) delivers user and signaling data to a specific UE in transport blocks. The size of a transport block is determined by the used modulation and coding scheme (MCS) and the transmission time length. A transport block can span across multiple subframes.

Signals used for cell search and synchronization are Narrowband Primary Synchronization Signals (NPSS) and Narrowband Secondary Synchronization Signals (NSSS). NPSS uses 11 subcarriers in the 6th subframe of every radio frame; NSSS carries the cell identity and information in the last subframe of every even radio frame. An eNB also transmits the System Information Block-NB (SIB-NB) through NPDSCH for extra system information not carried in MIB. For the uplink, there are two physical channels defined:

- Narrowband Physical Random Access Channel (NPRACH) is for eNB to receive random access preambles from UEs with single-tone 3.75 kHz subcarrier spacing.
- Narrowband Physical Uplink Shared Channel (NPUSCH) delivers user data and signals from a specific UE to the eNB using one of the two

formats. NPUSCH format 1 carries user data in single-tone or multitone tones. NPUSCH format 2 carries the acknowledgment (ACK) and negative ACK (NACK) of a Hybrid Automatic Repeat reQuest (HARQ) process.

Based on the radio resource specification, the NB-IoT system performs scheduling to determine the transport block size and the timing of NPDSCH and NPUSCH. The scheduling process takes into account received service data units and pre-allocated physical resources. NPDCCH carries scheduling parameters in DCI formats to the corresponding UE. Therefore, NB-IoT HARQ process, which completes a round of transport block transmission, sequentially transmits a DCI, transport blocks, and an ACK/NACK in the end. The HARQ in NB-IoT is both adaptive and asynchronous. There are five critical scheduling parameters.

- Resource assignments are the numbers of resource units assigned to the transport block.
- MCS index is to indicate the selected MCS. The combination with resource assignment determines the amount of data carried in a transport block.
- Repetition number is the number of transmission repetition applied to transport blocks. In NB-IoT, the radio uses repetition coding for error correction.
- Scheduling delay is the number of subframes to be postponed after DCI before transmitting transport blocks.
- Subcarrier indication specifies the number and location of used subcarriers affecting the duration of a transport block.

DCIs are defined in three formats: The DCI format N0 carries NPUSCH scheduling grants containing resource assignment, MCS index, repetition number, scheduling delay, and subcarrier indication. The DCI format N1 carries the scheduling control for NPDSCH and preambles for NPRACH contention free procedures, which contains scheduling parameters and the ACK resource index of the HARQ process. The DCI format N2 carries NPDSCH paging and

direct indication messages. A UE monitors predefined NPDCCH search spaces to detect the DCI messages. The NPDCCH period is the interval between two consecutive NPDCCH and is equal to $R_{\max} \times G$ subframes long. Allocation of NPDCCH essentially determines the frequency of scheduling opportunities (Huang et al. 2018).

The 3GPP TR 45.820 (2017) defines three levels of coverage enhancement (CE) achieved by repetition coding allowing up to 2048 repetitions. Normal CE0 is the level providing coverage similar to legacy General Packet Radio Service (GPRS). Robust CE1 provides about 10 dB improvement over CE0; Extreme CE2 provides about 20 dB improvement over CE0. The achieving target maximum coupling loss for CE0 to CE2 is 144, 154, and 164 dB, respectively.

To accomplish application data transmission procedures involving multiple transport blocks, NB-IoT supports three modes: legacy LTE service request, control plan optimization, and user plane optimization. The control plan optimization, which provides mobile-terminated (MT) and mobile-originated (MO) procedures, is the major lightweight protocol introduced below.

NB-IoT runs the MT procedure to send a data packet from the core network to a UE in the Radio Resource Control (RRC) idle state. After the paging process, the MT procedure sequentially schedules the following messages between the UE and an eNB:

1. Preamble. The UE sends a preamble signal to eNB to initiate the random access.
2. Random access response (RAR). RAR carries the response to the preamble and the grant of the next message.
3. RRC connection request. UE sends the connection request using uplink resources pre-allocated in RAR.
4. RRC connection setup. The eNB then schedules the setup information using the downlink to complete the random access process.
5. RRC connection setup complete. The uplink message from the UE establishes the RRC connection.

6. Downlink information transfer. After the RRC connection is established, the MT downlink data packet transmission begins.

For the MO procedure sending data from UE to core network, the first four messages are for the random access process as the same as the MT procedure. However, the RRC connection setup complete message carries the uplink data packet.

Key Applications

The mMTC is one of essential usage scenarios of the Fifth Generation (5G) systems (3GPP TR 38.913 2017) in the context of IoT. The NB-IoT is expected to be one of the extensively deployed technologies for mMTC. In this case, low data rate and low device cost applications ranging from smart city, smart building, environmental sensing, and consumer electronics can be well supported.

Cross-References

- [5G Wireless](#)
- [Machine-Type Communication](#)

References

- 3GPP TR 38913 (2017) Study on scenarios and requirements for next generation access technologies
- 3GPP TR 45820 (2017) Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT). TR 45.820, 3rd Generation Partnership Project (3GPP)
- 3GPP TS 23401 (2017) LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access
- 3GPP TS 36211 (2017) Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation
- 3GPP TS 36213 (2017) Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures
- Huang CW, Tseng SC, Lin P, Kawamoto Y (2018) Radio resource scheduling for narrowband Internet of Things systems: a performance study. IEEE Netw Mag

NB-IoT

- [Narrowband Internet of Things](#)

NDN

- [Mobility Management in NDN](#)

Near-Far Interference

- [Smart Device-to-Device Communication: Communication Establishment and Maintenance](#)

Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies

Phone Lin¹, Chi-Wei Tseng¹, Yi-Bing Lin², Shun-Ren Yang³, Chow-In Ko⁴, and Yichi Kawamoto⁵

¹Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

²Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

³Department of Computer Science and Information Engineering, National Tsing Hua University, Hsinchu, Taiwan, Republic of China

⁴Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan, Republic of China

⁵Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Synonyms

[D2D communications](#); [Decision-making](#); [Emergency medical services](#); [Healthcare](#); [Internet of Things \(IoT\)](#)

Definitions

The neighborhood-based ICT-driven support (NIS) system of emergency medical services (EMS) delivers efficient and interdisciplinary home care to elderly people in hyper-aged societies. The NIS system consists of five major components: (1) distributed NIS networks; (2) NIS slave node, an efficient sensing device; (3) NIS master node; (4) IoTtalk, an IoT device manager; and (5) NIS controller for EMS. Figure 1 illustrates an overview of the NIS system for elderly people.

Since many elderly people may not want to access new or advanced ICT technologies (e.g., smartphone or computer), the system should be designed as simple and automated as possible. Take the following community as an example, where each elderly's residence is no more than 50–100 m away from each other. We consider two types of elderly peoples. One is the cared elderly people who are older than 70 years old or are in poor physical conditions. The other is the caregiver elderly people who are younger (i.e., 60–65 years old) and healthier and is willing to use ICT products. The caregiver elderly people are responsible for regular care of cared elderly people.

Since the cared elderly people may not want to use ICT products, at the cared elderly people's place, a simple mobile device (i.e., the smartphone with Wi-Fi communication capability) is deployed, which is treated as an NIS slave node. The NIS slave node consists of applications which can report the health condition and collect data from the motion sensors at the cared elderly people's residence. The short-range device-to-device (D2D) communication technology is applied for relaying the data to the NIS controller via the NIS slave nodes. At the caregiver elderly people's residence, more advanced ICT technologies are deployed. A robot is treated as the NIS master node at the caregiver elderly people's residence. The robot is installed with intelligent functionalities and can monitor and have functionalities to work with the caregiver elderly people. In other words, the robot can "care of" the caregiver elderly people.

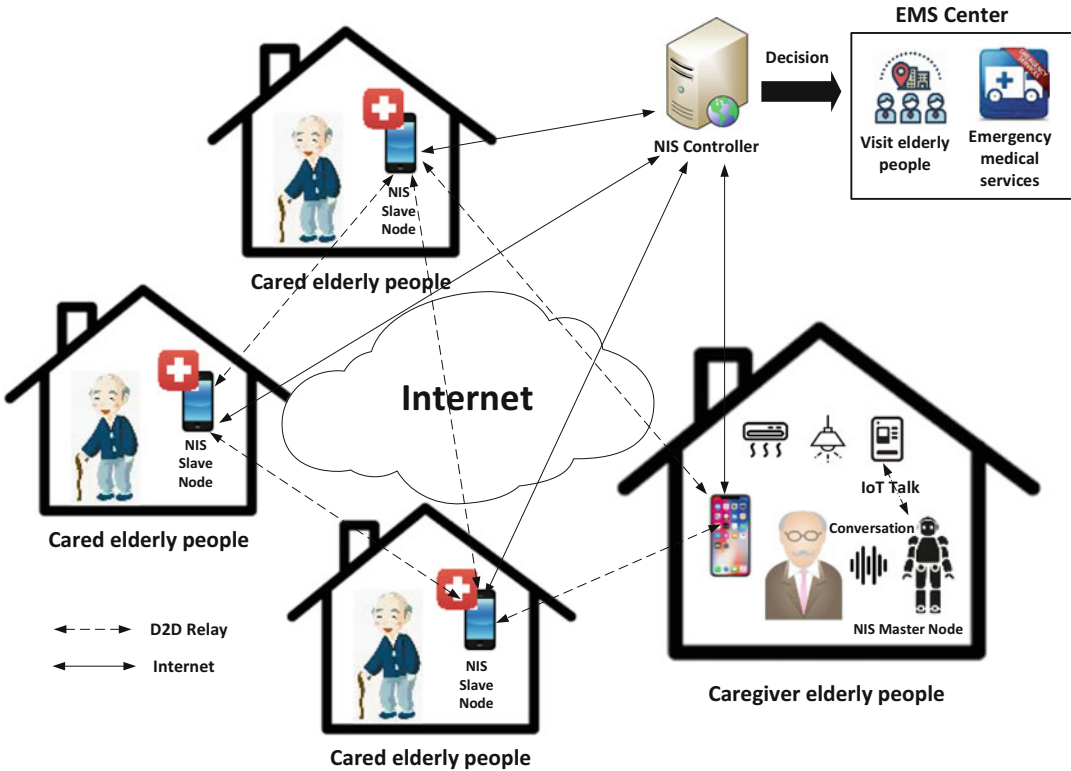
The concept of "community neighborhood" is that the elderly people (both cared and caregiver) who live in the same neighborhood can take care of each other. When abnormal activities from cared elderly people are detected, the caregiver elderly people are informed, and he could make a phone call or visit the cared elderly people personally for checking. The NIS master node has intelligence to assist caregiver elderly people to analyze and observe health condition of cared elderly people.

The health and medical monitoring data is transferred to the NIS controller for long-term and follow-up checking health condition of elderly people. An intelligent decision-maker in the NIS controller can determine the urgency of the cared elderly people. Then, the NIS controller could send an alert to EMS center for triggering any corresponding procedures to give aid to the cared elderly people. With the support of the NIS system, the emergency notification and response time of treatment for the elderly people can be drastically reduced, such that the situation of the lonely death could be avoided.

Historical Background

The world is aging rapidly. For example, the effect of aging is evident in both Japan and Taiwan. Japan is the most aged country in the world with more than 20% of its population over 65 years old. It is estimated that the number will reach 30.3% in 2025 and climb to 36.1% in 2040 ([Statistics Bureau](#)). Fourteen percent of Taiwanese people are anticipated to become over 65 years old in 2017, 20% in 2025, and more than 40% in 2060 ([Policy for long-term care: Ministry of Health and Welfare](#)).

Japan's policies are moving toward to having livelihood support services paid for something other than long-term insurance. In Japan, social security cost has exceeded 25% of the national income ([Ministry of Finance](#)). This could become the biggest cause of financial collapse, making conversion of existing long-term care policies come with a pressing issue. Furthermore, it is known that the number of lonely deaths (e.g.,



Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 1 An overview of NIS system

persons who live alone die without being noticed by anyone) of elderly people in Japan has been increasing in recent years (McDonald 2012).

Smart environments (e.g., smart home) often try to be aware of the user locations, situations, and lifestyles and try to provide proper services. For example, the system can automatically control home electrical appliances based on the home situation. The air conditioner may be turned on (or turned off) automatically based on the location and intention of the user. The intention of the user can be estimated based on the user's daily activity patterns. In addition, it is possible to avoid crisis by detecting the abnormal activities. In such circumstances, it is an important issue to build an efficient and effective community-based system in hyper-aged societies.

The neighborhood-based ICT-driven support (NIS) system for emergency medical services (EMS) fulfills the requirements of the elderly people care.

Foundations

A rise in the aging population increases the pressure on the social security and homecare industry to provide the necessary care. The IoT technology could be one solution to reduce the cost to resolve the issue.

The IoT allows different IoT devices connected to the Internet to work together without human invention. In the NIS system, the NIS master node can retrieve monitoring data from the NIS slave node. Analysis (e.g., machine learning technologies) on such information can help to know how cared elderly people behave in the daily life. Furthermore, if any behavior anomaly is detected, the caregiver can be informed first, and an alert could be sent to the EMS centers (e.g., 119 in Taiwan). Following that, a standard EMS service can be provided to the cared elderly people efficiently and effectively. Compared with the traditional EMS service, such a system can

speed up the whole procedure of the EMS service.

The NIS system consists of a robust communication infrastructure, simple NIS slave node for data collection, intelligent NIS master node, efficient devices management, and intelligent decision-maker for EMS response.

In the following, we illustrate the technologies that can be used in the five major components in the NIS system.

Distributed NIS Networks

The distributed NIS networks provide connectivity between the smartphones of the elderly residences by exploiting the Relay-by-Smartphone technology (Kang et al. 2007; Nishiyama et al. 2014, 2015a; Ito et al. 2013). Relay-by-Smartphone is a technology realizing device-to-device (D2D) communication by using the already available Wi-Fi on smartphones.

As shown in Fig. 2, it relays messages from device to device by employing the technologies of Wi-Fi Direct and delay-tolerant network (DTN) (Kawamoto et al. 2013; Takahashi et al. 2013, 2014; Oiyama et al. 2014).

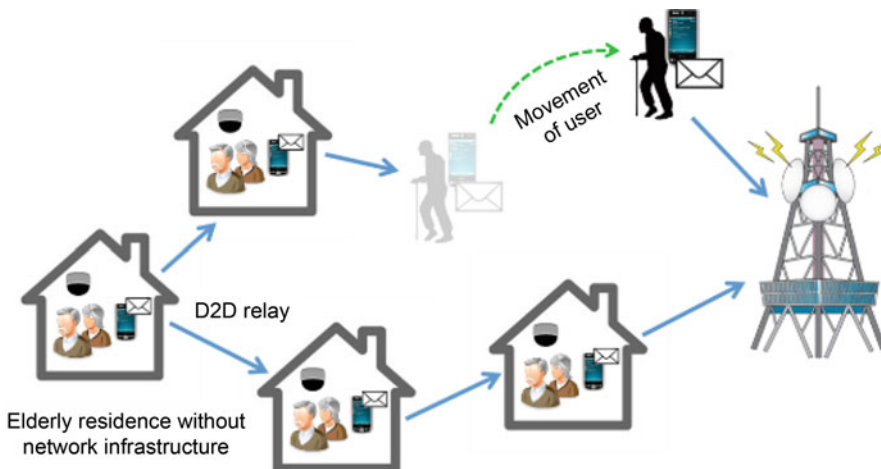
When the device can find a neighboring device around it, it can send the messages soon by using Wi-Fi Direct. On the other hand, when it is unable to find a neighboring device, it keeps the messages for a while and sends it after finding the

other device(s) or fixed infrastructures with DTN. It can be harnessed as a communication method in the locations where the network infrastructure such as cellular systems and open access points of the Internet are not available, for example, in rural areas, islands, mountainous areas, and so forth. The Relay-by-Smartphone technology is adopted to connect elderly residences in an isolated area for early detection of the occurrence of an emergency. In the hyper-aged societies, it is anticipated that the number of elderly residences without adequate communication tools increases, which causes many problems such as solitary deaths. Relay-by-Smartphone can provide a cheap yet effective communication scheme without fixed infrastructure because old and used smartphones are sufficient to install the application of the Relay-by-Smartphone.

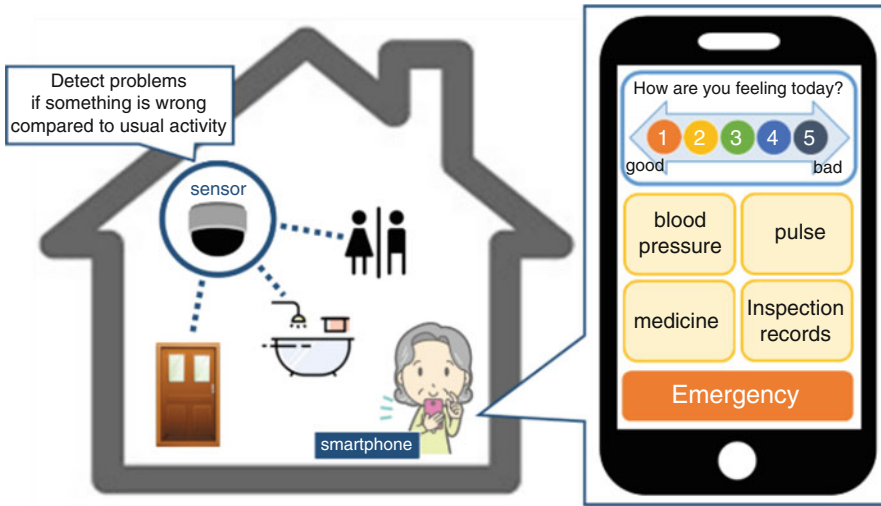
A network is constructed between the IoT sensors and smartphone of each elderly residence to collect and deliver the elderly people's health status and other monitored information over the Relay-by-Smartphone-based distributed network.

NIS Slave Node

Two applications are provided by the NIS slave node as shown in Fig. 3. One is for reporting the health conditions of the elderly people. It is operated on the smartphone as a simple application, for example, having buttons for reporting



Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 2 The overview of the constructed network based on Relay-by-Smartphone



Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 3 An overview of NIS slave node

the health condition on a five-point scale. It can be customized to include more inputs like blood pressure and pulse and whether they are taking medicines or not, according to the situation of each elderly people. Elderly people push the button by themselves at a certain time of everyday. By using this, the system can monitor the change of the health condition of elderly people and analyze what kind of procedure is required for them. In addition, the application can be used for reporting the occurrence of an emergency situation. By pushing the emergency button, they easily notice the situation to neighborhoods and emergency units.

The other application is using the motion sensor at the elderly residence. The motion sensor monitors the daily motion of elderly people such as opening and closing of the door of restroom, entrance of the house, and so forth and sends the information via the constructed network in Topic1. The system learns the usual motion of each elderly people and detects problems if something is wrong compared to usual activity based on the learning and reported information from the sensor.

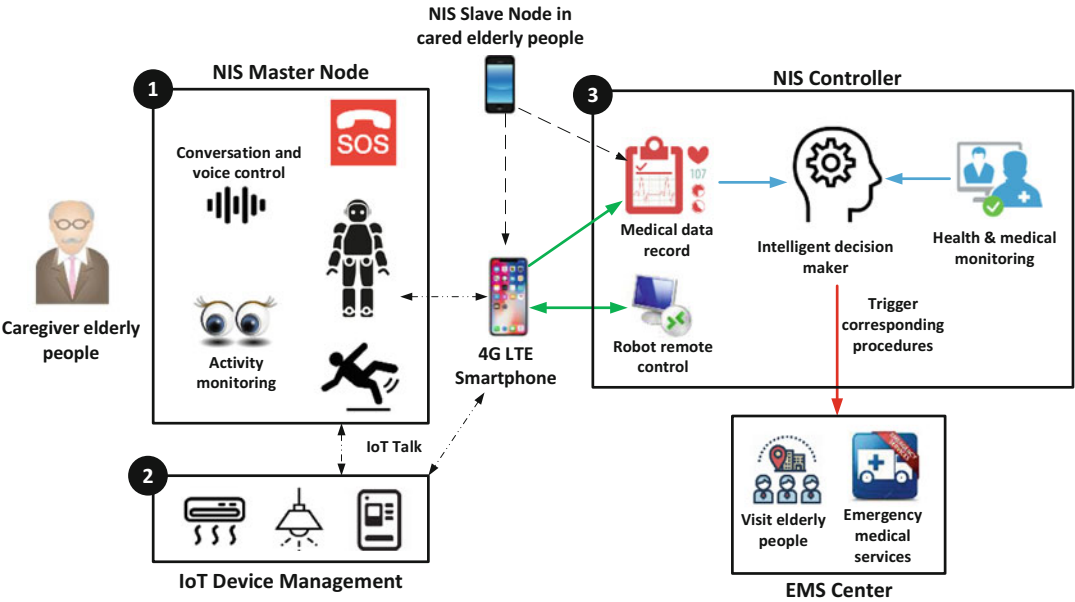
NIS Master Node

The NIS master node performs two major functions. First, it collects the data from the NIS slave

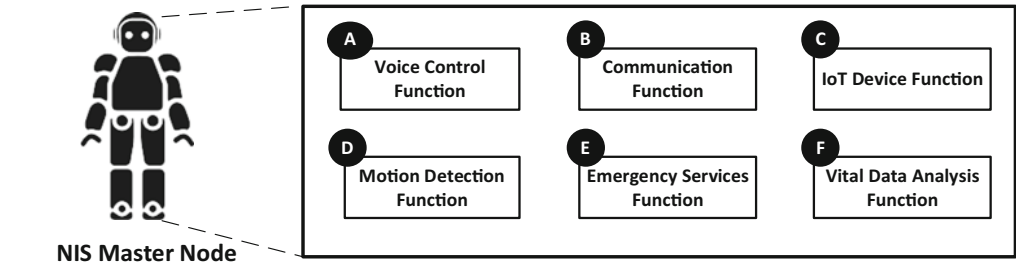
node at the cared elderly people's place and analyzes the monitoring data to build the behavioral model for the cared elderly people, which is used to detect any abnormal behavior of the cared elderly people. If such an event is detected, it will inform the caregiver elderly people to check the status of the cared elderly people and send out an alert to the EMS centers if necessary.

Second, the NIS master node monitors the activities of the caregiver elderly people and can have simple interaction with the caregiver elderly people. For example, if it is too hot indoor, the NIS master node can turn on the air conditioner for the caregiver elderly people. Figure 4 illustrates the system architecture of the NIS master node, NIS controller, IoT device management, and IoTalk. The related works that have been proposed on robot technology for daily healthcare of elderly or disabled people can be found in Nishiyama et al. (2015b), Kim et al. (2009), Pung et al. (2009), and Kudo (2012).

In the NIS master node, an autonomous robot can execute homecare tasks: reminding elderly people to take medicines, measuring the physiological information, and monitoring the activity of the elderly people. The functional module of the NIS master node is shown in Fig. 5. Some commercial robots (e.g., ZenboTM



Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 4 Framework of NIS master node, NIS controller, and IoT device management



Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 5 Functions of the NIS master node

(Ahn et al. 2017)) are based on Android Software Development Kit (SDK) (<https://zenbo.asus.com/developer/documents/Zenbo-SDK-Getting-Started/Getting-Started>), and the system developer can implement the features with application programming interfaces (APIs) provided by the robot manufacturer.

To summarize, an NIS master node (see Fig. 4 (1)) serves as an intelligent robot, which consists six main function modules detailed below:

(A) Voice control interface serves as the interface for the users to control the behavior of the robot. The voice recognition techniques are

applied to let the robot understand the intentions from the caregiver elderly people and execute the corresponding instructions.

- (B) Communication function consists of Wi-Fi and Bluetooth communication functions. The robot can communicate with the IoT devices or the NIS slave node directly.
- (C) IoT device function controls the IoT devices based on the IoTtalk API (Lin et al. 2015). This function enables interaction between the devices and sensors in the house.
- (D) Motion detection function monitors and detects possible abnormal behaviors by checking the activities of elderly daily living (ADL) (Crepeau 2003; Suzuki et al. 2006).

- (E) Emergency service function enables the caregiver elderly people to be able to issue an order to the NIS master node to trigger the emergency medical service when cared elderly people are in emergency situation.
- (F) Vital data analysis function performs vital data sensing at home, which includes blood pressure (using sphygmomanometer with Bluetooth), blood sugar level (using blood glucose meter with Bluetooth), body temperature, and so on. Analysis (e.g., machine learning technologies) on such information can help to know how cared elderly people behave in the daily life.

IoT Device Management for NIS System

The NIS system adopts the IoTtalk technology (also known as IoT EasyConnect system) (Lin et al. 2015) to manage IoT devices. An IoT device is characterized by its “features” (e.g., temperature, vibration, and display) that are manipulated by the network applications. If a network application handles the individual device features independently, a software module can be implemented for each device feature, and the network application can be simply constructed by including these brick-like device feature modules. Based on the concept of device feature, the applications and interactions for homecare equipment are easily created.

The IoTtalk technology is adopted to manage the environment devices at the caregiver elderly’s home (see Fig. 4 (2)). For instance, when the sensor detects an elderly people’s step length has become short, the IoTtalk may control other facilities at home to initiate appropriate response (e.g., switching on the light, recommending the use of a cane through home audio/visual system, and so forth).

Intelligent Decision-Maker at NIS Controller

An intelligent decision-maker is installed at NIS controller (see Fig. 4 (3)), which uses the AI technology (e.g., machine learning) to analyze the health and medical data collected from the elderly people. According to the data processing by the intelligent decision-maker, we may classify three

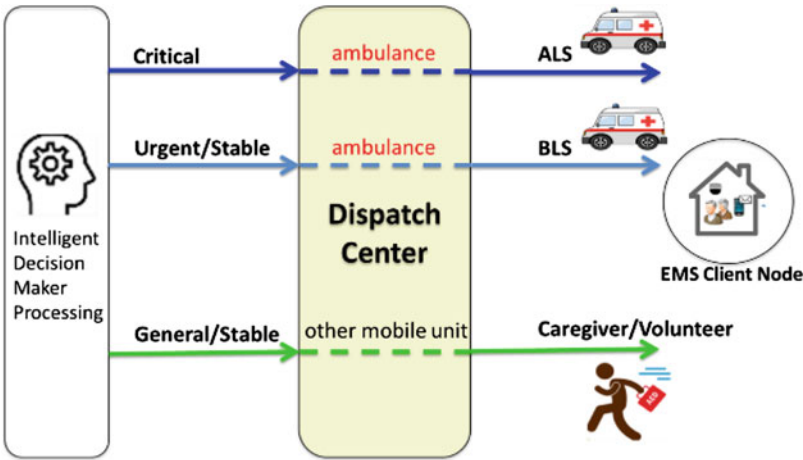
categories of abnormal situations: (1) critical, (2) urgent but stable, and (3) general and stable. An optimal processing algorithm uses the collected sensor inputs to determine whether any abnormal situation occurs.

Key Applications

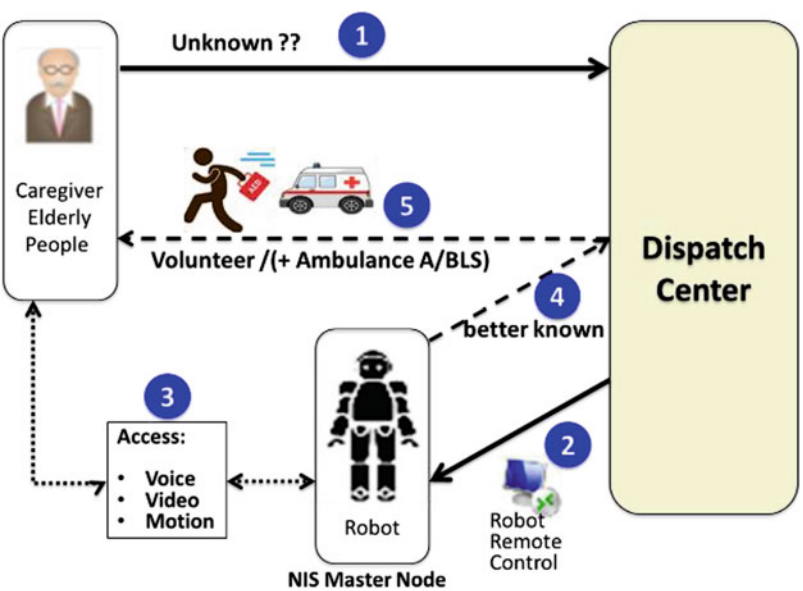
The key application where the NIS system can be applied is the emergency medical services (EMS) also known as ambulance services or paramedic services. Figure 6 shows a general case of triggering the corresponding EMS management. The dispatch center receives each type of alarm and then reacts different responses properly according to the corresponding category (Lin et al. 2015). The NIS has two major types of EMS responses: (1) by EMS ambulance and (2) by other mobile unit other than EMS ambulance (Ng et al. 2017; Ringh et al. 2015; <http://www.pulsepoint.org/pulsepoint-respond/>; https://www.scdf.gov.sg/content/scdf_internet/en/community-and-volunteers/mobile_phone_technology.html). EMS ambulance only responds for two categories of situation: (1) critical and (2) urgent but stable.

There are often two levels of EMS ambulance, ALS (advanced life support) and BLS (basic life support). Only ALS team is capable of providing intravenous resuscitating medication, endotracheal intubation, and manual defibrillation. ALS ambulance would be reserved for the “critical” cases, which often imply patients with unstable vital signs, and BLS ambulance would be utilized for the “urgent but stable” cases (Lin et al. 2015). To reduce the false alarms and to save ambulance resources for the mostly critical patients and avoid ambulance misuse, for the “general and stable” cases, instead of utilizing EMS ambulance, other designated mobile units including the caregiver, the neighborhood volunteers, or the other community first responders such as guards running by bike, feet, or motorbike would visit the target elderly first and then arrange the proper transport if they really need to seek the hospital care (Lin et al. 2015; <http://www.pulsepoint.org/pulsepoint-respond/>; <https://www.pulsepoint.org/pulsepoint-respond/>).

Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 6 Triggering of corresponding EMS managements at the NIS controller



Neighborhood-Based ICT-Driven Support (NIS) System of Emergency Medical Services (EMS) for Elderly People in Hyper-Aged Societies, Fig. 7 Triggering of corresponding EMS managements to the caregiver elderly people



scdf.gov.sg/content/scdf_internet/en/community-and-volunteers/mobile_phone_technology.html).

Figure 7 illustrates the model of triggering the corresponding EMS managements specific to the caregiver elderly people at NIS master node. The details are described as follows:

- Step 1. When the caregiver elderly people have a problem, the abnormal situation may not be well known because of its server node and caregiver character not being cared as client node by other elderlies.
- Step 2. The dispatch center may actively activate the robot remote control to make it access the caregiver.

- Step 3. In NIS server node, the robot remote control could access caregiver’s situation through voice, video, or motion detection functions to get the situation category.
- Step 4. The remote and automatic assessments by robot of caregiver’s situation would be transmitted back to the dispatch center to gather better information for prompt EMS response.
- Step 5. The designated rapid response volunteers for assistance to caregiver would be soon activated to NIS server node to check caregiver’s condition and to provide helps. Before their arrival if the robot assessments showed urgent or critical status, the EMS ambulance would

be sent out concurrently with the rapid volunteer response (Lin et al. 2015; <http://www.pulsepoint.org/pulsepoint-respond/>; https://www.scdf.gov.sg/content/scdf_internet/en/community-and-volunteers/mobile_phone_technology.html).

References

- Ahn HS, Lee MH, Broadbent E, Macdonald BA (2017) Gathering healthcare service robot requirements from young people's perceptions of an older care robot. 2017 First IEEE International Conference on Robotic Computing (IRC). <https://doi.org/10.1109/irc.2017.48>
- Crepeau EB (2003) Willard & Speckman's occupational therapy, Lippincott Williams and Wilkins, Philadelphia, United States
- Ito M, Nishiyama H, Kato N (2013) A novel routing method for improving message delivery delay in hybrid DTN-MANET networks. 2013 IEEE Global Communications Conference (GLOBECOM). <https://doi.org/10.1109/glocom.2013.6831050>
- Kang K, Lee J, Choi H (2007) Instant notification service for ubiquitous personal care in healthcare application. 2007 International Conference on Convergence Information Technology (ICCIT 2007). <https://doi.org/10.1109/iccit.2007.4420466>
- Kawamoto Y, Nishiyama H, Kato N (2013) Toward terminal-to-terminal communication networks: a hybrid MANET and DTN approach. 2013 IEEE 18th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD). <https://doi.org/10.1109/camad.2013.6708122>
- Kim M, Kim S, Park S et al.(2009) Service robot for the elderly. IEEE Robot Autom Mag 16:34–45. <https://doi.org/10.1109/mra.2008.931636>
- Kudo M (2012) Robot-assisted healthcare support for an aging society. 2012 Annual SRH Global Conference. <https://doi.org/10.1109/srhi.2012.36>
- Lin Y-B, Lin Y-W, Chih C-Y et al.(2015) EasyConnect: a management system for IoT devices and its applications for interactive design and art. IEEE Internet Things J 2:551–561. <https://doi.org/10.1109/jiot.2015.2423286>
- McDonald M (2012) In Japan, lonely deaths in society's margins. [Online]. Available: <http://rendezvous.blogs.nytimes.com/2012/03/25/injapan-lonely-deaths-in-societys-margins>
- Ministry of Finance website. Retrieved 20 Aug 2016, from http://warp.ndl.go.jp/info:ndljp/pid/11178065/www.mof.go.jp/budget/fiscal_condition/related_data/sy014/sy014q.htm
- Ng Y, Leong S, Ong M (2017) The role of dispatch in resuscitation. Singap Med J 58:449–452. <https://doi.org/10.11622/smedj.2017059>
- Nishiyama H, Ito M, Kato N (2014) Relay-by-smartphone: realizing multihop device-to-device communications. IEEE Commun Mag 52:56–65. <https://doi.org/10.1109/mcom.2014.6807947>
- Nishiyama H, Ngo T, Oiyama S, Kato N (2015a) Relay by smart device: innovative communications for efficient information sharing among vehicles and pedestrians. IEEE Veh Technol Mag 10:54–62. <https://doi.org/10.1109/mvt.2015.2481558>
- Nishiyama H, Takahashi A, Kato N et al.(2015b) Dynamic replication and forwarding control based on node surroundings in cooperative delay-tolerant networks. IEEE Trans Parallel Distrib Syst 26:2711–2719. <https://doi.org/10.1109/tpds.2014.2359889>
- Oiyama S, Nishiyama H, Kato N (2014) A partially centralized messaging control scheme using star topology in delay and disruption tolerant networks. International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2014). <https://doi.org/10.1109/spects.2014.6879995>
- Policy for long-term care: Ministry of health and welfare. Retrieved 23 May 2018., from <http://www.ndc.gov.tw/en/cp.aspx?n=2E5DCB04C64512CC&s=002ABF0E676F4DB5>
- Pung H, Gu T, Xue W et al.(2009) Context-aware middleware for pervasive elderly homecare. IEEE J Sel Areas Commun 27:510–524. <https://doi.org/10.1109/jsac.2009.090513>
- Ringh M, Rosenqvist M, Hollenberg J et al.(2015) Mobile-phone dispatch of laypersons for CPR in out-of-hospital cardiac arrest. Resuscitation 96:2. <https://doi.org/10.1016/j.resuscitation.2015.09.008>
- Statistics Bureau website. Retrieved 23 May 2018., from <http://www.stat.go.jp/data/topics/topi901.html>
- Suzuki R, Ogawa M, Otake S et al.(2006) Rhythm of daily living and detection of atypical days for elderly people living alone as determined with a monitoring system. J Telemed Telecare 12:208–214. <https://doi.org/10.1258/13576330677488780>
- Takahashi A, Nishiyama H, Kato N (2013) Fairness issue in message delivery in delay- and Disruption-Tolerant Networks for disaster areas. 2013 International Conference on Computing, Networking and Communications (ICNC). <https://doi.org/10.1109/icnc.2013.6504207>
- Takahashi A, Nishiyama H, Kato N et al.(2014) Replication control for ensuring reliability of convergecast message delivery in infrastructure-aided DTNs. IEEE Trans Veh Technol 63:3223–3231. <https://doi.org/10.1109/tvt.2014.2299288>

Network Big Data

► Big Network Data

Network Capacity

- [Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space](#)
- [Capacity of Wireless Ad Hoc Networks](#)

Network Coding

- [Network Coding-Aware Routing in Wireless Ad Hoc Networks](#)

Network Coding-Aware Routing

- [Network Coding-Aware Routing in Wireless Ad Hoc Networks](#)

Network Coding-Aware Routing in Wireless Ad Hoc Networks

Yan Yan and Baoxian Zhang
University of Chinese Academy of Sciences,
Beijing, China

Synonyms

[Network coding](#); [Network coding-aware routing](#)

Definition

Network coding-aware wireless routing is a routing paradigm that takes into account network coding gains and broadcast benefit at wireless nodes in route selection for supporting efficient data delivery in wireless networks to improve the network performance.

Historical Background

Coding-aware routing protocols can be classified into two categories: centralized protocols and distributed protocols.

Centralized protocols depend on centralized linear programming formulations in which a central controller is responsible for making the routing decisions for all requests by using global state information. Some typical work in this aspect is as follows. Sengupta et al. (2007) proposed a linear-programming-based interflow coding-aware routing protocol. The authors derived a theoretical formulation for computing the network throughput on any wireless topology with any pattern of concurrent unicast sessions by making routing decisions with the awareness of potential network coding opportunities at different nodes. Ni et al. (2006) proposed a coding-aware routing protocol using linear programming to improve the network performance by reducing the expected number of coded transmissions for successful exchange of packets via interflow network coding. Although these centralized coding-aware routing protocols can improve the network performance in terms of resource utilization and throughput, their centralized characteristic leads to the scalability problem. Moreover, centralized protocols have the issue of single point of failure and the central controller could be a bottleneck of the whole network. Therefore, they are not suitable for being employed in dynamic multihop wireless ad hoc networks.

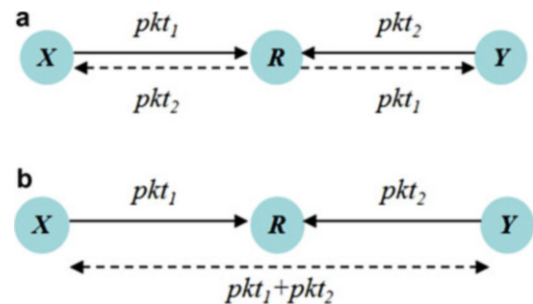
In distributed coding-aware routing protocols, nodes in the network independently make localized routing decisions based on the network state information each of them locally collects/maintains. In the literature, there are two types of distributed protocols: source routing protocols and hop-by-hop routing protocols. In source routing, the source node of a flow locally computes an end-to-end path from the source node to the destination node with more coding gains by using certain global state information (including global network topology, state information of each link in the network, and paths taken by other communication

requests). In hop-by-hop routing, each routing decision is made based only on the limited local state information that the packet holder collects/maintains. As the result, hop-by-hop routing has low complexity/overhead and is thus suitable for being deployed in dynamic wireless ad hoc networks. In recent years, several hop-by-hop coding-aware routing protocols have been proposed. In Wu et al. (2007), to support efficient interflow network coding, Wu et al. designed a Markovian routing metric, which is defined as the cost of sending a packet across a link and its value depends on the link where the packet comes from. This metric can work with an existing routing protocol to choose paths with more network-coding opportunities. Le et al. (2008) proposed an on-demand distributed coding-aware unicast routing protocol (DCAR) for selecting routes with more potential coding gains to improve the network throughput by introducing a coding-aware routing metric that jointly considers potential coding opportunities, congestion levels, and other related factors. Yan et al. (2008) designed a rate-adaptive coding-aware multiple path routing protocol (RCR) for wireless multihop networks. The main design objective is to improve the network performance via traffic splitting for maximizing the coding opportunities in the wireless networks. Yan et al. (2016) considered the interplay between transmit power, transmit rate, and network coding gain in a wireless ad hoc network and they proposed a distributed network coding aware power/rate control protocol, which allow each node to adaptively adjust its transmit power and data rate such that spatial-temporal bandwidth resource in the network is fully utilized via using efficient coding-aware transmissions, whenever applicable. All the above four protocols were designed based on the traditional best-path routing protocol, and none of them considered impact of the lossy characteristic of wireless channel on the efficiency of coded packet transmissions. To respect such impact, CORE (Yan et al. 2010) is designed to improve the link-level data delivery reliability and network throughput and it combines hop-by-hop opportunistic forwarding and localized

interflow network coding by exploiting the lossy characteristic of wireless channel. A common feature of all the above five protocols is that they focus on interflow network coding-based routing for improved network performance. Some other work (e.g., Chung et al. 2012; Seferoglu et al. 2011; Khreishah et al. 2013) had also considered hybrid approaches that exploit both interflow network coding and intraflow network coding for lossy wireless ad hoc networks where interflow network coding considers encoding packets from different flows and intraflow network coding considers encoding packets from the same flow.

Foundations

Recently, wireless network coding has been proposed as an effective paradigm to achieve high throughput performance of wireless ad hoc networks. Wireless network coding can obtain network coding gains from the inherent broadcast nature of wireless medium in the wireless networks by mixing multiple packets in a single transmission. Figure 1 illustrates the basic idea of network coding in a wireless network, which was first reported in Sengupta et al. (2007). In Fig. 1, node X and node Y want to exchange packets via a common intermediate node R , X sends packet pkt_1 to Y and Y sends packet pkt_2 to X . After node R receives both packets pkt_1 and pkt_2 , it can create a new xor-coded packet " $pkt_1 \text{ xor } pkt_2$ " and then broadcast it to the air. Upon the receipt of the coded packet, both X and Y can decode their



Network Coding-Aware Routing in Wireless Ad Hoc Networks, Fig. 1 A simple example illustrating how wireless coding improves the throughput of wireless networks. (a) No coding and (b) coding

interested packet by using the packet received and the packet information that it locally stores. The number of total transmissions in this example is thus reduced from four to three.

With the assistance of wireless network coding, coding-aware routing protocols combine localized network coding and route selection to significantly improve the network throughput performance and also reduce the number of transmissions in wireless networks. The objective of coding-aware routing is to enable efficient communications in wireless networks by reducing the number of packet transmissions by utilizing network coding opportunities and also improve the network performance such as network throughput, packet delivery delay, etc. Therefore, implementation of coding aware routing in wireless networks requires to consider both network coding gains and routing cost (e.g., delay, throughput, and packet loss) in traditional routing protocols, in order to achieve improved network performance effectively.

The basic idea behind coding-aware routing can be illustrated by the following example. Consider the example network in Fig. 2. Suppose there are two flows: one from node *A* to node *E* and the other from node *F* to node *A*, and further suppose there exist no loss over wireless links. The traditional shortest paths for these two flows are $A \rightarrow B \rightarrow C \rightarrow E$ and $F \rightarrow D \rightarrow B \rightarrow A$,

respectively. In this case, we can see that there exists no coding opportunity if both *A* and *F* decide to forward packet along their corresponding shortest paths. However, if *F* can forward its packets via *E* to *A*, then the path from *F* to *A* will be detoured onto $F \rightarrow E \rightarrow C \rightarrow B \rightarrow A$. Although this path has one more extra hop than the corresponding shortest path, it creates coding opportunities at nodes *C* and *B*. The total number of packet transmissions in this case will be reduced by the coding-aware forwarding strategy. The network throughput can also be improved as a result.

Key Applications

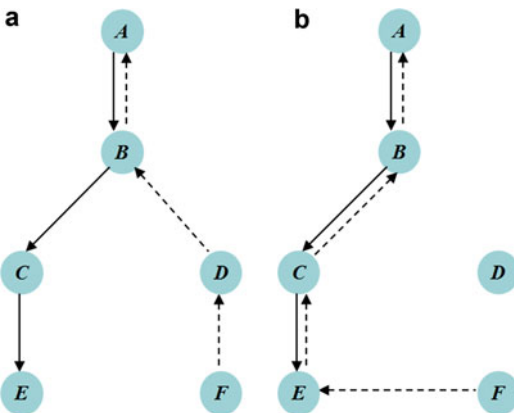
Network coding-aware wireless routing is beneficial in wireless multihop networks where network coding opportunities exist and efficient wireless routing is required to improve the network resource utilization efficiency.

Cross-References

- [Network Coding](#)
- [Routing](#)

References

- Chung K, Chou Y-C, Liao W (2012) CAOR: coding-aware opportunistic routing in wireless ad hoc networks. In: Proceedings of IEEE international conference on communications (ICC), Ottawa, Canada, June 2012, pp 136–140
- Khreishah A, Khalil I, Wu J (2013) Low complexity and provably efficient algorithm for joint inter and intrasession network coding in wireless networks. *IEEE Trans Parallel Distrib Syst* 24(10):2015–2024
- Le J, Lui JCS, Chiu DM (2008) DCAR: distributed coding-aware routing in wireless networks. In: Proceedings of IEEE ICDCS 2008, Beijing, China, June 2008, pp 462–469
- Ni B, Santhapuri N, Zhong Z, Nelakuditi S (2006) Routing with opportunistically coded exchanges in wireless mesh networks. In: Proceedings of WiMesh'06, Reston, VA, Sept 2006, pp 157–159
- Seferoglu H, Markopoulou A, Ramakrishnan KK (2011) I²NC: intra- and inter-session network coding for unicast flows in wireless networks. In: Proceedings of the



Network Coding-Aware Routing in Wireless Ad Hoc Networks, Fig. 2 An example illustrating the idea of coding-aware routing. (a) Shortest path routing and (b) coding aware routing

- IEEE INFOCOM 2011, Apr 2011, Shanghai, China, pp 1035–1043
- Sengupta S, Rayanchu S, Banerjee S (2007) An analysis of wireless network coding for unicast sessions: the case for coding-aware routing. In: Proceedings of IEEE INFOCOM'07, Anchorage, AK, May 2007, pp 1028–1036
- Wu Y, Das SM, Chandra R (2007) Routing with a Markovian metric to promote local mixing. In: Proceedings of IEEE INFOCOM 2007, Anchorage, AK, May 2007, pp 2381–2385
- Yan Y, Zhao Z, Zhang B, Mouftah HT, Ma J (2008) Rate-adaptive coding-aware multiple path routing for wireless mesh networks. In: Proceedings of IEEE Globecom 2008, New Orleans, USA, Nov 2008, pp 543–547
- Yan Y, Zhang B, Zheng J, Ma J (2010) CORE: a coding-aware opportunistic routing mechanism for wireless mesh networks. *IEEE Wirel Commun* 17(3):96–103
- Yan Y, Zhang B, Yao Z (2016) Space-time efficient network coding for wireless multi-hop networks. *Comput Commun* 73(part A):144–156

Network Effects

Zehui Xiong¹, Dusit Niyato², and Ping Wang^{1,3}

¹School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore, Singapore

²School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore, Singapore

³Department of Information and Communication Engineering, Tongji University, Shanghai, China

Synonyms

Demand-side economies of scale; Network externalities

Definitions

Network effect has been defined as the value or surplus that an agent derives from the goods/service when the number of other agents consuming the same kind of goods/service changes, which could be positive or negative.

Historical Background

Network economics, inherently, is an interdisciplinary research area between economics and networks. Economics is the social science that analyzes the production, distribution, and consumption of goods and services. As for networks, the focus is on connectivity, which could be referred to as market network with many network industries, social networks that permeate our society and daily life, or the communication networks (e.g., Internet) in engineering area. Equipped with economic theory, the economic features of networks can be analyzed. Undoubtedly, network modeling and analysis in economics is more than a fad. The ubiquity of networked interactions in economic settings clearly shows that network economics analysis should continue to grow rapidly (Bramoulle et al. 2016).

There is a fundamental property of networks, i.e., the fact that they exhibit network externalities due to complementarity of network structure (Economides 1996). The seemingly simple phenomenon has brought a great influence to the majority of networking progress, including network formulation, network competition or coalition, and network segmentation, hence to a large extent determines market strategies and outcomes in network markets. Considering the value of one node without looking into its activity in the network or without incorporating the role it can play in growing the network due to network externality is myopic. Therefore as the core issue of network economics, the study for network externality is highly significant and necessary for providing a guide toward understanding the network itself. Besides, network resource allocation policies are a major issue of networking progress, while pricing is a powerful countermeasure for allocation issues in most times, which inevitably demands our attention.

Network externality has been defined as the surplus that an agent derives from the service when the number of other agents consuming the same kind of service changes, which could be positive or negative (Shapiro and Varian 2013; Liebowitz and Margolis 1994; Katz and

Shapiro 1985; Liebowitz and Margolis 1998; Farrell and Saloner 1985). Fax machine and telephones are two obvious positive examples, which become increasingly valuable and hence more attractive to one user if there are more users adopting it, similar to Facebook and Twitter in social networks. Strictly speaking, *network effects* should not properly be called *network externalities*. Unfortunately, the term externality has been used somewhat carelessly in many economic literatures which cause some ambiguity.

Early literatures denote the network effects and network externalities as the same, with the meaning or description when the value of a product to one user depends on the total number of adopters (for discussion see (Katz and Shapiro 1985; Shapiro and Varian 2013)). Another portion of literature denotes the network externalities are caused by network effects. Specifically, *network externalities* refer to the economic concept of the externalities entailed in network effects. When the owner of a network or participants is able to internalize such network effects, they are no longer externalities (Liebowitz and Margolis 1998, 1994). Recently, some papers define network effects as one special case of positive network externalities; also they extend the meaning of network externality (Huang and Gao 2013), which is more realistic. In their definition, network externalities reflect the situation where each user's utility is affected by other users' decisions. After all, any economic effect is an externality if not internalized since the externality is more general (Farrell and Klemperer 2007). But it has not been adopted by many scholars.

Putting aside definitional concerns, this chapter merely focuses on considering and discussing the extra value owing to the network effects and gives a relative comprehensive review on the topic from different perspective. *Network effects* are implied as positive in most times. For brief, this chapter uses the term "*network effects*" to replace positive network effects in subsequent discussion. Furthermore, this chapter follows the most economic literature's idea and predefines *network effects* as the phenomenon when a good or service becomes more valuable when more

people use it, which is denoted as positive network externality (Shapiro and Varian 2013). The chapter may use these two terms interchangeably for the same meaning.

The "*network effect*" is often highlighted in social and economic viewpoints (Tucker 2008); however until recently, network effects have been discussed and applied in communication networks. As for data communication networks, many papers simply assume the network effects are not dominant like congestion or interference effects hence easily ignored without discussion. They obviously leave out some situations with the presence of significant network effects, where the attention should be paid enough, which is exactly the motivation for the chapter. For example, the authors in Blackburn et al. (2013) studied the users' calling behaviors concerning the usage of cellular service such as data, SMS message, and voice calls, which affects economics of cellular provider with the analysis of real telecommunication service data. Furthermore, the real data is leveraged to validate the existence of underlying network effects. Additionally, the authors in Blackburn et al. (2013) provides the pioneering motivation for planning and developing the pricing mechanisms for future work. As one of the key issues of network economics (Chiu and Ng 2011), network effects exist in many connected communication network platforms such as Internet, mobile, and ad hoc network and peer-to-peer communication networks (Easley and Kleinberg 2010).

Preliminaries of Pricing Network Effects

This section briefly introduces and describes basic definitions in network economics, strengths or weakness, and mathematical models that can be referred to later in the sequel.

Definitions and Interpretation

Network externalities can be either positive or negative. The negative network ones have been extensively investigated in many existing research works (Shy 2011; Chiu and Ng

2011), such as the congestion and interference in wired and wireless networks in Gibbens et al. (2000) and MacKie-Mason and Varian (1995), respectively. However, as for positive externalities or network effects, there are a limited number of literatures especially on communications. This chapter does not focus our attention on negative externalities in this chapter, yet the situation – positive network externalities appear together with negative ones – is considered, which was investigated probably for the first time in Chao (1996) and mainly takes place in communication networks (Johari and Kumar 2010; Gong et al. 2015; Zhang et al. 2016a, 2017a).

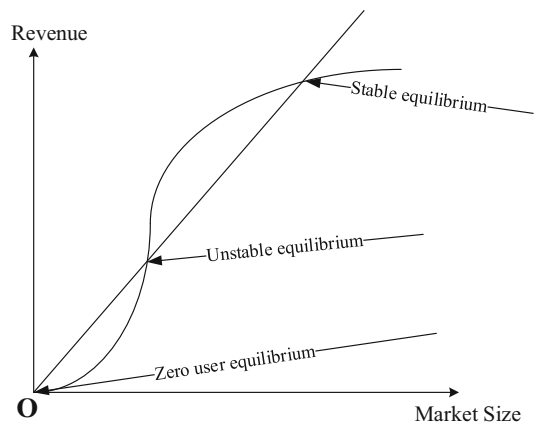
In the literature, the network effects are mainly categorized into three types. Direct network effects or same-side network effects have been defined as those generated through a direct physical effect of the number of consumptions on the value of service or good. For example, in peer-to-peer communication networks, more users increase the value of the network by giving more destination options. Indirect network effects are usually the cases where complementary goods provider gains more value as the number of users of a good increases. When the interaction is reciprocal, this chapter calls it two-sided network effects and cross-side or cross-group network effects, which are a special case of indirect network effects (Rochet and Tirole 2006). Again take the peer-to-peer communication networks, for example, the more existing users enable the peering network to function at lower cost and higher capacity as well as service quality, which in turn benefit the users themselves.

Local (peer-to-peer) network effects or social network effects consider the localized effects of users rather than the global network effects (direct or indirect), where each user is influenced directly by the decisions of only a typically small subset of other users (Sundararajan 2005, 2007). For example, Zune players can share music with people around, so the utility obtained by a Zune player directly depends on the number of its friends sharing the music on Zune. Another more direct and relevant case is instant messaging, the utility of one user is critically determined

by the number of its friends who use the same instant messenger. Generally speaking, the local network effects mainly focus on the structure of an underlying social network.

Benefits and Problems

As discussed in the last section, the pricing mechanism is mainly used for revenue maximization by service or goods providers as a methodology of allocation policies, which are conducive to appeal to users. The growth of the adopter base is represented as one of the most significant metrics to reflect success. Generally speaking, traditional price increases with quantity demanded of the service or products due to economic equilibrium; hence the usage-based pricing is frequently used in network services. However due to the presence of network effects, another important consideration – size of network – which has a direct relation with the value obtained by users requires our attention. In Fig. 1, assume the cost is fixed for selling the service/products in a market network; the total cost of all users is denoted as a straight line with the increase of consumers/users. Due to network effects the real value of service/products is represented as a sigmoidal curve, which has three cross-points (market equilibria) with former straight line, the middle equilibrium is not stable because any change regarding market size causes the network to converge to one of another two equilibria.

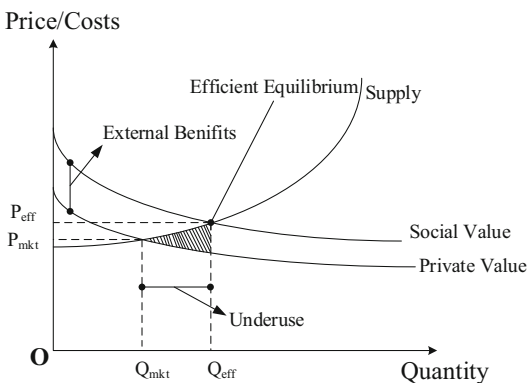


Network Effects, Fig. 1 Market equilibria with network effects

The similar observation comes from many other studies involving network effect (Chiu and Ng 2011).

From another specific perspective in Fig. 2, it is intuitively observed that there are more details on external benefits for the sake of network effects. Here external benefits are denoted as difference of social value and private value. The social benefit (or revenue) includes not only the consumers' own benefits of getting the service/products (marginal benefits) but also the external benefits. Thus, the social marginal revenue is larger than the private marginal revenue, as shown in Fig. 2. Each of value curves has an intersection with supply (private costs) curve, which denotes the corresponding equilibrium.

Moreover, the optimum demand from social level is Q_{eff} that denotes the efficient equilibria in terms of quantity, which is larger than the normal market demand Q_{mkt} from the level of private profit maximization. When a product has external benefits due to network effects, the market equilibrium is lower and there remains underused resources that cause deadweight loss (shadow areas), so how does the provider handle the situations? Obviously, the corresponding pricing strategies should be designed with the consideration of extra benefits caused by network effects. Intuitively, the provider should take measures to shift up the demand (private value) curve and then achieve the higher efficient equilibrium. One common solution is Pigouvian subsidy which



Network Effects, Fig. 2 Problems (market failures) without taking into consideration of network effects

encourages people to consume while guarantees that private value plus subsidy equals to social value. Generally, the network effect plays an important role in improving the performance of a pricing strategy.

Application Scenario in Wireless Networks with Network Effects

The pricing schemes are designed to offer profitable business to the service providers, as well as to create favorable services for the users, which is a useful tool for allocation issues (Huang and Gao 2013). Since users typically share limited resources in data communication networks, congestion effect from physical domain on users' behavior is common and has been extensively investigated for pricing strategies by many works (MacKie-Mason and Varian 1995; Blumrosen and Dobzinski 2007; Sengupta and Chatterjee 2009; Hande et al. 2010). For example, if an Internet service provider becomes oversubscribed, its subscribers suffer from the congestion externality due to limited bandwidth and radio resources. However, very few works of network effects take homophily phenomenon into consideration, i.e., network effects. The social aspect of mobile networking is an emerging paradigm for network design and optimization (Gong et al. 2017b). In Brown and Reingen (1987), the authors found that the information obtained from social tie connections will influence in decision-making. Inspired by Brown and Reingen (1987), the social group utility maximization framework was proposed in Gong et al. (2017a) and Chen et al. (2016), which captures the impact of users' diverse social ties that are subject to diverse social relationships. The authors in Blackburn et al. (2013) showed the evidence of network effect in communication service using the real data analytic and quantified such effect using a simple metric. The finding of Blackburn et al. (2013) provides the motivation for planning pricing mechanisms with network effect in the real markets.

Recently, the network effect has been considered jointly with service pricing from the economic perspective. For example, in the pioneering work (Candogan et al. 2012), the authors investigated both the uniform pricing and the discriminatory pricing of the service provider in the presence of network effects. In Makhdoumi et al. (2017) and Swapna et al. (2012), the authors discussed the dynamic pricing strategy of divisible social goods with network effects. However, the authors investigated the user behaviors only in social networks without considering the congestion in wireless network. Therefore, the model introduced in Makhdoumi et al. (2017) has a major limitation as it cannot be applied into the wireless network environment, in which the radio resource is scarce and congestion can frequently happen. Subsequently, the authors in Gong et al. (2017b) formulated the interaction between a wireless service provider and mobile users as a two-stage Stackelberg game, based on the model presented in Candogan et al. (2012). In the upper Stage I of the formulated game, the service provider acting as the leader determines the price for users. Then, the users acting as the followers simultaneously decide on the data usage level maximizing their individual utilities given the price in the lower Stage II. The uniform pricing scheme was adopted, where the impacts of network effects and congestion were jointly analyzed. In Zhang et al. (2016b, 2017a), the authors studied the similar problem to that in Gong et al. (2017b), where the discriminatory pricing strategy was applied by the provider. The authors in Zhang et al. (2016a) further investigated a market where multiple wireless service providers exist and focused on the competition among the providers. In Xiong et al. (2017b), the authors studied the sequential dynamic pricing scheme where the user's behavior are subject to both the network effects and congestion effects. Specifically, the provider can utilize its ability to modify its strategy in response to the observed history. Similarly, the authors in Zhang et al. (2017b) analyzed the impacts of network effects on cloud computing resource allocation. The authors in Nie et al. (2017) exploited the network effects to design the incentive mecha-

nism for engaging more participants in the mobile crowdsensing platform.

Another socially aware communication network is the sponsored content. Sponsored content enables a content provider to pay a wireless network service provider (SP), and thereby mobile users can access contents from the content provider (CP) through network services from the service provider with lower charge. Thus, more users want to access the contents which potentially generate more profit gain to the content provider. The authors in Wang et al. (2017) focused on the service-selection process among the mobile users as an evolutionary population game and demonstrated how global network effects and sponsoring helps improve the SP's revenue and the mobile user's experience. The authors in Xiong et al. (2017a) conducted economic analysis on sponsored content in the presence of social network effects, where the interactions among the SP, CP, and MUs are modeled as a three-stage game. Under the similar model, the authors in Xiong et al. (2018b) further investigated the sequential competition and cooperation between the CP and SP. Lately, the authors in Xiong et al. (2018a, submitted) proposed a joint sponsored and edge caching content model under noncooperative game framework, where the mobile users can access sponsored content using 4G link and edge caching content via Wi-Fi link. Therein, the mutual interplay between the edge caching CP and sponsored CP is the main concern.

Conclusions

"Network effects" is an interesting phenomenon whereby "more connected users and more benefits," which originally comes from network economics. At the same time, pricing has been widely investigated for service or product's provider in terms of revenue maximization all this time. This chapter focuses on the "network effects"-based pricing from analytical and practical perspective. The work lies in the area of pricing as well as revenue maximization in the presence of network effects, with an emphasis

on communication services. Meanwhile, this chapter has implemented with the tutorial parts with respect to basic definitions and models on pricing and network effects. Furthermore, this chapter has presented several actual application scenarios on pricing network effects. The hope of this chapter is that the novel insights from economic and sociological viewpoints will be helpful to researchers from telecommunication's area regarding pricing of emerging socially aware service with network effects and enable the engineers to apply the economic strategies from one subject to another. Many questions remain open and may be ripe for subsequent theoretical and even practical investigations in the next few years.

Key Applications

Mobile or wireless social networks where the network effects and congestion effects both exist, or some emerging communication applications such as crowdsensing, user provided network, and sponsored content platform.

References

- Blackburn J, Stanojevic R, Erramilli V, Iamnitchi A, Pagiannaki K (2013) Last call for the buffet: economics of cellular networks. In: Proceedings of the 19th ACM annual international conference on Mobile computing & networking, pp 111–122
- Blumrosen L, Dobzinski S (2007) Welfare maximization in congestion games. *IEEE J Sel Areas Commun (JSAC)* 25(6):1224–1236
- Bramoulle Y, Galeotti A, Rogers B (2016) The Oxford handbook of the economics of networks. Oxford University Press, New York
- Brown JJ, Reingen PH (1987) Social ties and word-of-mouth referral behavior. *J Consum Res* 14(3):350–362
- Candogan O, Bimpikis K, Ozdaglar A (2012) Optimal pricing in networks with externalities. *Oper Res* 60(4):883–905
- Chao SY (1996) Positive and negative externality effects on product pricing and capacity planning. PhD thesis, Citeseer
- Chen X, Gong X, Yang L, Zhang J (2016) Exploiting social tie structure for cooperative wireless networking: a social group utility maximization framework. *IEEE/ACM Trans Netw (ToN)* 24(6):3593–3606
- Chiu DM, Ng WY (2011) Exploring network economics. arXiv preprint arXiv:1106.1282
- Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, New York
- Economides N (1996) The economics of networks. *Int J Ind Organ* 14(6):673–699
- Farrell J, Klemperer P (2007) Coordination and lock-in: competition with switching costs and network effects. In: Armstrong M, Porter R (eds) Handbook of industrial organization, vol 3. Elsevier, North Holland, pp 1967–2072
- Farrell J, Saloner G (1985) Standardization, compatibility, and innovation. *The RAND J Econ* 16:70–83
- Gibbens R, Mason R, Steinberg R (2000) Internet service classes under competition. *IEEE J Sel Areas Commun (JSAC)* 18(12):2490–2498
- Gong X, Duan L, Chen X (2015) When network effect meets congestion effect: leveraging social services for wireless services. In: Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing, pp 147–156
- Gong X, Chen X, Xing K, Shin DH, Zhang M, Zhang J (2017a) From social group utility maximization to personalized location privacy in mobile networks. *IEEE/ACM Trans Netw (ToN)* 25(3):1703–1716
- Gong X, Duan L, Chen X, Zhang J (2017b) When social network effect meets congestion effect in wireless networks: data usage equilibrium and optimal pricing. *IEEE J Sel Areas Commun (JSAC)* 35(2):449–462
- Hande P, Chiang M, Calderbank R, Zhang J (2010) Pricing under constraints in access networks: revenue maximization and congestion management. In: Proceedings of IEEE INFOCOM, San Diego
- Huang J, Gao L (2013) Wireless network pricing. *Synth Lect Commun Netw* 6(2):1–176
- Johari R, Kumar S (2010) Congestible services and network effects. *EC* 10:93–94
- Katz ML, Shapiro C (1985) Network externalities, competition, and compatibility. *Am Econ Rev* 75(3):424–440
- Liebowitz SJ, Margolis SE (1994) Network externality: an uncommon tragedy. *J Econ Perspect* 8(2):133–150
- Liebowitz SJ, Margolis SE (1998) Network externalities (effects). *New Palgrave's Dictionary of Economics and the Law*, New York
- MacKie-Mason JK, Varian HR (1995) Pricing congestible network resources. *IEEE J Sel Areas Commun (JSAC)* 13(7):1141–1149
- Makhdoumi A, Malekian A, Ozdaglar A (2017) Optimal pricing policy of network goods. In: Annual allerton conference on communication, control, and computing (allerton), Monticello
- Nie J, Xiong Z, Niyato D, Wang P, Luo J (2017) A socially-aware incentive mechanism for mobile

crowdsensing service market. arXiv preprint arXiv:171101050

Rochet JC, Tirole J (2006) Two-sided markets: a progress report. *RAND J Econ* 37(3):645–667

Sengupta S, Chatterjee M (2009) An economic framework for dynamic spectrum access and service pricing. *IEEE/ACM Trans Netw (ToN)* 17(4):1200–1213

Shapiro C, Varian HR (2013) *Information rules: a strategic guide to the network economy*. Harvard Business Press, Boston

Shy O (2011) A short survey of network economics. *Rev Ind Organ* 38(2):119–149

Sundararajan A (2005) Local network effects and network structure. Information systems working papers series, Stern School of Business, New York University

Sundararajan A (2007) Local network effects and complex network structure. *BE J Theor Econ* 7(1):1–35

Swapna B, Eryilmaz A, Shroff NB (2012) Dynamic pricing strategies for social networks in the presence of externalities. In: *Information theory and applications workshop (ITA)*, San Diego, pp 374–381

Tucker C (2008) Identifying formal and informal influence in technology adoption with network externalities. *Manag Sci* 54(12):2024–2038

Wang W, Xiong Z, Niyato D, Wang P (2017) A hierarchical game with strategy evolution for mobile sponsored content/service markets. In: *Proceedings of IEEE GLOBECOM*, Singapore

Xiong Z, Feng S, Niyato D, Wang P, Zhang Y (2017a) Economic analysis of network effects on sponsored content: a hierarchical game theoretic approach. In: *Proceedings of IEEE GLOBECOM*, Singapore

Xiong Z, Feng S, Niyato D, Wang P, Han Z (2017b) Network effect-based sequential dynamic pricing for mobile social data market. In: *Proceedings of IEEE GLOBECOM*, Singapore

Xiong Z, Feng S, Niyato D, Wang P, Leshem A, Han Z (2018a, submitted) Joint sponsored and edge caching content service market: a game-theoretic approach

Xiong Z, Feng S, Niyato D, Wang P, Zhang Y (2018b) Competition and cooperation analysis for data sponsored market: a network effects model. In: *Proceedings of IEEE WCNC*, Barcelona

Zhang M, Yang L, Gong X, Zhang J (2016a) Impact of network effect and congestion effect on price competition among wireless service providers. In: *IEEE annual conference on information science and systems (CISS)*, Princeton, NJ

Zhang X, Guo L, Li M, Fang Y (2016b) Social-enabled data offloading via mobile participation-a game-theoretical approach. In: *Proceedings of IEEE GLOBECOM*, Washington, DC

Zhang X, Guo L, Li M, Fang Y (2017a) Motivating human-enabled mobile participation for data offloading. *IEEE Trans Mobile Comput* 17:1623–1636

Zhang Y, Xiong Z, Niyato D, Wang P, Jin J (2017b) A game-theoretic analysis of complementarity, substitutability and externalities in cloud services. In: *Proceedings of IEEE GLOBECOM*, Singapore

Network Externalities

► Network Effects

Network Function Virtualization

► Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization

Network Information Theory

Xiugang Wu

Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Synonyms

[Multiuser information theory](#)

Definitions

Network information theory is the extension of the classical information theory for point-to-point communications to networks of nodes exchanging information.

Historical Background

In 1948, Claude Shannon, the father of information theory, published his revolutionary paper “A Mathematical Theory of Communication” (Shannon 1948). In the paper, the most significant result is arguably given by the channel coding theorem, which answers one of the most fundamental questions in information theory, namely,

what is the ultimate transmission rate of communication. Roughly speaking, the channel coding theorem says that there always exists a maximum achievable rate for a channel, called the channel capacity, below which the reliable communication can be implemented while above which the reliable communication is impossible.

While Shannon's classical information theory addresses point-to-point channels and has had profound impact on modern communication systems, network information theory (El Gamal and Kim 2011) considers communication networks containing multiple senders and receivers, e.g., computer networks and satellite networks, and its main goal is to find the fundamental limits in network communications and the optimal coding schemes to achieve these limits. The presence of multiple senders and receivers in the network settings brings in many new elements to the communication problem, such as interference, cooperation, and feedback, but also makes the problems difficult to solve. To date, there is as yet no unified theory of network information flow, although such a complete theory is expected to have wide implications for today's communication networks, especially the currently ubiquitous wireless communication networks.

Problems in Network Information Theory

Given that the solution to the general problem is out of reach, various special problems in network information theory have been studied extensively over the past few decades, including distributed source coding, multiple-access channel, broadcast channel, interference channel, relay channel, etc. In distributed source coding, multiple (two or more) sources are separately encoded at their corresponding rates, and the descriptions are communicated over noiseless links to a decoder who wishes to recover all the sources either losslessly or within some distortion; the problem is to determine the set of simultaneously achievable description rate vectors, i.e., the rate vectors at which the sources can be recovered in a lossless or lossy sense. While distributed source

coding remains an open problem in general, it has been solved in the lossless case, and the achievable scheme is now known as Slepian-Wolf coding (Slepian and Wolf 1973). In a multiple-access channel, multiple senders send information to a common receiver, and the problem is to determine the capacity region, i.e., the set of rate vectors at which the senders can reliably communicate to the receiver simultaneously. The multiple-access channel has been completely solved, and one can achieve its capacity either by successive cancellation decoding or by simultaneous decoding (Ahlsweide 1973). The broadcast channel describes the scenario where there is one sender and multiple receivers. While the capacity region for general broadcast channels is still unknown, it has been determined for some special classes, for example, the degraded broadcast channels where the capacity is achieved by superposition coding (Cover 1972). The interference channel (in the simplest setting) has two senders and two receivers, with the signals of the two senders interfering with each other. The capacity region of the interference channel is not known in general except for several special cases such as the strong interference case (Sato 1981); a prominent achievable scheme for the interference channel is the Han-Kobayashi coding scheme which employs rate splitting and superposition coding (Han and Kobayashi 1981). The relay channel models a communication scenario where there is a relay node that can help the information transmission between the source and destination. While the capacity of the relay channel is generally unknown, substantial advances have been made in Cover and Gamal (1979), where a general upper bound on the capacity, now known as the cutset bound, together with two fundamental relay schemes, namely, decode-forward and compress-forward, is developed.

Recent Progress

Motivated by the increasing development of wireless networks, the past decade has witnessed significant progress in network information theory, on both the achievability and converse

sides. On the achievability side, a plethora of new coding schemes have been developed for network problems, ranging from the extension and unification of decode-forward and compress-forward to the case of multiple relays (Xie and Kumar 2005; Kramer et al. 2005; Lim et al. 2011; Wu and Xie 2013, 2014), to schemes based on interference alignment (Cadambe and Jafar 2008) and distributed MIMO (Ozgur et al. 2007), to lattice-based techniques (Nazer and Gastpar 2011), and to strategies inspired by network coding and linear deterministic models (Ahlsvede et al. 2000; Avestimehr et al. 2011). On the converse side, proof techniques that build on geometric and functional analysis have been introduced very recently, which lead to the solutions of several long-standing open problems in the field. For example, an open question regarding the capacity of the relay channel has been solved via a geometric approach in Wu et al. (2016), and the “missing corner point” in interference channel has been settled by using transportation-information inequalities in Polyanskiy and Wu (2016).

References

- Ahlsvede R (1973) Multi-way communication channels. In: Second international symposium on information theory: Tsahkadsor, Armenia, USSR, 2–8 Sept 1971
- Ahlsvede R, Cai N, Li SY, Yeung RW (2000) Network information flow. *IEEE Trans Inf Theory* 46(4): 1204–1216
- Avestimehr AS, Diggavi SN, David N (2011) Wireless network information flow: a deterministic approach. *IEEE Trans Inf Theory* 57(4):1872–1905
- Cadambe VR, Jafar SA (2008) Interference alignment and degrees of freedom of the k -user interference channel. *IEEE Trans Inf Theory* 54(8):3425–3441
- Cover T (1972) Broadcast channels. *IEEE Trans Inf Theory* 18(1):2–14
- Cover T, Gamal AE (1979) Capacity theorems for the relay channel. *IEEE Trans Inf Theory* 25(5):572–584
- El Gamal A, Kim YH (2011) Network information theory. Cambridge University Press
- Han T, Kobayashi K (1981) A new achievable rate region for the interference channel. *IEEE Trans Inf Theory* 27(1):49–60
- Kramer G, Gastpar M, Gupta P (2005) Cooperative strategies and capacity theorems for relay networks. *IEEE Trans Inf Theory* 51(9):3037–3063
- Lim SH, Kim YH, El Gamal A, Chung SY (2011) Noisy network coding. *IEEE Trans Inf Theory* 57(5):3132–3152
- Nazer B, Gastpar M (2011) Compute-and-forward: harnessing interference through structured codes. *IEEE Trans Inf Theory* 57(10):6463–6486
- Ozgur A, L  v  que O, David N (2007) Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *IEEE Trans Inf Theory* 53(10):3549–3572
- Polyanskiy Y, Wu Y (2016) Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Trans Inf Theory* 62(7):3992–4002
- Sato H (1981) The capacity of the gaussian interference channel under strong interference (corresp.). *IEEE Trans Inf Theory* 27(6):786–788
- Shannon CE (1948) A mathematical theory of communication, part I, part II. *Bell Syst Tech J* 27:623–656
- Slepian D, Wolf J (1973) Noiseless coding of correlated information sources. *IEEE Trans Inf Theory* 19(4):471–480
- Wu X, Xie LL (2013) On the optimal compressions in the compress-and-forward relay schemes. *IEEE Trans Inf Theory* 59(5):2613–2628
- Wu X, Xie LL (2014) A unified relay framework with both df and cf relay nodes. *IEEE Trans Inf Theory* 60(1):586–604
- Wu X, Barnes LP,   zg  r A (2016) Cover’s open problem: the capacity of the relay channel. In: 2016 54th annual allerton conference on communication, control, and computing, Allerton. IEEE, pp 148–155
- Xie LL, Kumar PR (2005) An achievable rate for the multiple-level relay channel. *IEEE Trans Inf Theory* 51(4):1348–1358

Network Maximum Throughput

- [Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space](#)

Network MIMO

Jun Zhang
Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Sai Kung, Hong Kong

Synonyms

[Cooperative MIMO](#); [Virtual MIMO](#)

Definition

Network MIMO is a family of cooperative multiantenna transmission/reception techniques for wireless communications. With network MIMO, multiple wireless access points share user data and channel state information via high-capacity backhaul links and perform joint precoding to transmit to multiple user devices in the downlink and jointly decode the received signals for different users in the uplink.

Historical Background

Point-to-point multiantenna transmission and reception techniques, known as *single-user MIMO*, were developed in 1990s to support high-speed wireless communications. Being able to greatly increase throughput via utilizing spatial domain resources without bandwidth expansion, they formed key enablers for broadband wireless communications systems, including wireless local area networks (IEEE 802.11n and IEEE 802.11ac) and cellular networks (WiMAX and LTE). The great success of MIMO drove efforts to further exploit its potential, and various theoretical studies and practical signal processing techniques were then developed for *multiuser MIMO* systems. Since 2001, the concept of *network MIMO* arose, mainly in the form of base station cooperation in cellular networks. It was proposed to reap the technical maturity of traditional MIMO techniques and also to overcome the fundamental limitation in cellular networks, i.e., the intercell interference caused by co-channel communications in multiple cells.

By sharing user data and channel state information among multiple nodes, network MIMO transforms an interference limited system into a virtual multiuser MIMO system. Initial studies investigated the application of various multiuser MIMO processing techniques in the network MIMO setting. For example, Shamai and Zaidel (2001) applied dirty paper coding for base station cooperation in the downlink, which

is the optimal scheme for multiuser MIMO. In Zhang and Dai (2004), both dirty paper coding approach and several practical joint transmission schemes were investigated for downlink base station cooperation, with a per base station power constraint.

The performance gains of network MIMO techniques, especially their ability to mitigate intercell interference, attracted great attention from the industry. Bell Labs, which demonstrated the first MIMO system, also pioneered in the investigation on network MIMO. In a series of papers (Foschini et al. 2005, 2006; Venkatesan et al. 2007), they confirmed that network MIMO is capable to significantly increase throughput under realistic propagation and operational conditions. In particular, the name of *network MIMO* first appeared in Venkatesan et al. (2007).

Around 2008, as the LTE standard was close to completion, the importance of interference management in the new LTE network was well recognized. Various interference coordination and cooperation techniques were tested, and simple coordination schemes were included in the LTE standard, e.g., fractional frequency reuse (FFR) was supported by LTE Release 8, and enhanced intercell interference coordination (eICIC) was supported by LTE Release 10. To further enhance the performance, a work item on coordinated multipoint (CoMP) transmission/reception was formed in 3GPP LTE-Advanced, which unified several network MIMO techniques. In 2011, CoMP was included as a key feature in LTE Release 11 (Irmer et al. 2011).

Foundations

Network MIMO builds upon traditional MIMO techniques, including single-user MIMO and multiuser MIMO, and it can overcome their limitations. It can effectively mitigate intercell interference and thus is attractive for cellular networks.

By deploying multiple antennas at the communication nodes, MIMO techniques can effectively

increase the capacity of wireless communications systems via spatial signal processing, without bandwidth expansion. The capacity of a single-user MIMO system grows linearly with the minimum number of transmit and receive antennas. Due to the space limitation, however, each mobile device can only have a small number of antennas, normally one or two, which thus bounds its capacity gain. Multiuser MIMO, where a base station communicates with multiple mobile users simultaneously, provides an opportunity to boost the total throughput even with single-antenna users. Nevertheless, for multiuser MIMO serving a large number of users, the sum capacity is restricted by the number of antennas at each base station. By cooperating multiple base stations, network MIMO essentially creates a giant virtual base station, and thus, it can increase the available spatial resources and overcome the limitation of multiuser MIMO.

Another difficulty in deploying traditional MIMO techniques in practical systems is the effect of intercell interference. The cellular network by design is interference-limited, and the transmission in each cell acts as interference to other cells. The situation gets worse as universal frequency reuse is advocated in modern wireless networks. While the problem of intercell interference is inherent to cellular systems, its effect on MIMO is more prominent as each neighboring base station antenna element acts as a unique interfering source, thereby making it difficult for the mobile user to estimate and suppress them. Studies (Catreux et al. 2000; Blum 2003) have shown that the capacity gains promised by MIMO techniques are degraded severely in the multicell environment. Different from conventional methods to suppress intercell interference, network MIMO attacks this problem from a fundamentally different perspective. With joint processing among multiple base stations, network MIMO turns otherwise harmful interference into useful signals, i.e., it eliminates intercell interference. This is especially important for cell edge users, who previously suffer from severe intercell interference and now can get a performance boost by receiving

from/transmitting to multiple cooperative base stations.

Network MIMO techniques are attractive due to their significant performance gains. Their practical implementation, however, faces several key challenges, including the complexity of the cooperative processing, the requirement of a high-capacity backhaul network, the high signaling overhead, and channel state information (CSI) acquisition. Joint processing among base stations significantly increases the computational complexity. Network MIMO requires data sharing among multiple base stations, which calls for high-capacity backhaul links and significantly increases the signaling overhead. In order to support joint processing, CSI of all the users to all the cooperating base stations is needed, which causes significant training and feedback overhead. Great efforts have been spent to address these issues, and following are some potential techniques and solutions.

- *Low-complexity joint precoding and decoding:* To reduce the computational complexity of joint processing in network MIMO, linear schemes are preferred, e.g., zero-forcing precoding (Huh et al. 2012; Zhang 2010). Such schemes have been developed for multiuser MIMO, and they can be modified for network MIMO systems. One major difference is that per base station power constraint should be considered. Another approach to reduce complexity is via decentralized precoding, which allows each base station to locally process its received signal (Ho et al. 2011).
- *Clustered cooperation:* Clustering is an effective means to reduce the processing complexity, the amount of data sharing between base stations, as well as the required CSI. Base stations within each cluster can eliminate intracluster interference among users via joint precoding/decoding, which only requires local data and CSI sharing within the cluster. To improve the cluster edge performance, limited intercluster coordination methods can be applied to suppress interference for cluster edge users (Zhang et al. 2009).

- *Finite-capacity backhaul*: Theoretical performance gains of network MIMO are obtained by assuming high-capacity backhaul links. In practice, backhaul links are with finite capacity, which will affect the achievable data rate. One method to improve the performance with finite-capacity backhaul links is to adapt the cooperation scheme based on channel realization (Marsch and Fettweis 2008). Novel compression techniques should also be developed to deal with limited backhaul capacity, e.g., via joint signal and CSI compression (Kang et al. 2014).
- *CSI acquisition*: Accurate CSI of each user to all the cooperating base stations plays a critical role in network MIMO. Taking downlink transmission as an example, such information can be obtained via downlink training and uplink feedback. For downlink training, it is important to determine the amount of pilot symbols needed, which depends not only on the network size, but also on the backhaul capacity (Hoydis et al. 2011). To feedback estimated CSI to the base stations, each user can apply adaptive limited feedback to improve the performance with a given feedback bit budget, which partitions the available feedback bits between the desired and interfering channels (Bhagavatula and Heath 2011).

Key Applications

Network MIMO can be applied in wireless networks where distributed access points are connected via high-capacity backhaul links. It helps to suppress interference and improve the network throughput, as well as cell edge throughput.

Cross-References

- [Base Station Cooperations under Imperfect Conditions](#)
- [Coordinated Multipoint Transmission and Reception](#)
- [Network Slicing-Enabled Green C-RAN](#)

References

- Bhagavatula R, Heath RW Jr (2011) Adaptive limited feedback for sum-rate maximizing beamforming in cooperative multicell systems. *IEEE Trans Signal Process* 59(2):800–811
- Blum RS (2003) MIMO capacity with interference. *IEEE J Select Areas Commun* 21(5):793–801
- Catreux S, Driessen PF, Greenstein LJ (2000) Simulation results for an interference-limited multiple-input multiple-output cellular system. *IEEE Commun Lett* 4(11):334–336
- Foschini GJ, Huang H, Karakayali K et al (2005) The value of coherent base station coordination. In: *Proceedings of the conference on information sciences and systems (CISS)*, Mar 2005. John Hopkins University
- Foschini GJ, Karakayali K, Valenzuela RA (2006) Enormous spectral efficiency of isolated multiple antenna links emerges in a coordinated cellular network. *IEEE Proc Commun* 153(4):548–555
- Ho W, Quek TQS, Sun S, Heath RW Jr (2011) Decentralized precoding for multicell MIMO downlink. *IEEE Trans Wirel Commun* 10(6):1798–1809
- Hoydis J, Kobayashi M, Debbah M (2011) Optimal channel training in uplink network MIMO systems. *IEEE Trans Signal Process* 59(6):2824–2833
- Huh H, Tulino AM, Caire G (2012) Network MIMO with linear zero-forcing beamforming: large system analysis impact of channel estimation and reduced-complexity scheduling. *IEEE Trans Inf Theory* 58(5):2911–2934
- Irmr R, Droste H, Marsch P et al (2011) Coordinated multipoint: concepts performance and field trial results. *IEEE Commun Mag* 49(2):102–111
- Kang J, Simeone O, Kang J et al (2014) Joint signal and channel state information compression for the backhaul of uplink network MIMO systems. *IEEE Trans Wirel Commun* 13(3):1555–1567
- Marsch P, Fettweis G (2008) On base station cooperation schemes for downlink network MIMO under a constrained backhaul. In: *Proceedings of the IEEE global telecommunication conference*, New Orleans, Nov–Dec 2008
- Shamai S, Zaidel BM (2001) Enhancing the cellular downlink capacity via co-processing at the transmitting end. In: *Proceedings of the IEEE vehicular technology conference*, Rhodes, May 2001
- Venkatesan S, Lozano A, Valenzuela R (2007) Network MIMO: overcoming intercell interference in indoor wireless systems. In: *The forty-first asilomar conference on signals, systems and computers*, Nov 2007, Pacific Grove
- Zhang R (2010) Cooperative multi-cell block diagonalization with per-base-station power constraints. *IEEE J Sel Areas Commun* 28(9):1435–1445
- Zhang H, Dai H (2004) Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks. *Eur J Wirel Commun Netw* 2004:222–235
- Zhang J, Chen R, Andrews JG et al (2009) Networked MIMO with clustered linear precoding. *IEEE Trans Wirel Commun* 8(4):1910–1921

Network Privacy Supervision and Privacy Preserving

► [Network Privacy Surveillance and Privacy Protection](#)

Network Privacy Surveillance and Privacy Protection

Changgen Peng
Guizhou Provincial Key Laboratory of Public
Big Data, Guizhou University, Guiyang, China

Synonyms

[Network privacy supervision and privacy preserving](#)

Definitions

Network privacy surveillance and protection issues include the protection of individual sensitive information in collection, storage, analysis processes and their application for administration and economic interest. It is also concerned with the balance between privacy protection and usability.

Historical Backgrounds

Privacy surveillance and protection are fundamental topics arising to be our increasing concerns in the era of the Internet (Lyon and Zureik 1996; Ashrafi and Kuilboer 2018). The governments in many countries have started to collect individual private information from public administration purpose even before such information was regarded as private (Warren and Brandeis 1890). With new legislation efforts government is gaining more control over private life of public (Parliament of Canada 2003; Young 2010; Meece 2012). In addition, new

technologies such as Cloud Computation, Big Data, Mobile Internet and Internet of Things, even though they have benefited our daily life magically, enable governments and corporations to monitor public technically. Similarly, Video surveillance cameras, biometrics, domestic drones, face recognition technology, RFID chips, and stingray tracking devices also become the basic privacy collecting and surveillance technologies (American Civil Liberties Union 2013), facilitating governments and corporations in the collection of more and more private information, and storage and application in of those information for administration and economics reasons. Privacy becomes prone to be abused, resulting in violation of individual rights, informational inequality, informational injustice and discrimination, and encroachment on moral autonomy. Therefore, it is extremely important to protect privacy, find appropriate balance between information utilization and privacy protection.

Privacy Definition and Identification Technology

As privacy is closely related to individual preferences, independence and the entire society, it's not easy to define privacy precisely. The professor of law (Westin 2003) believes that privacy is something focusing on *individuals*. This understanding is considered too broad to define privacy accurately. But from the perspective of the application, privacy is mostly depended on the characteristics of the application scenario. According to the difference of service objects, privacy definitions can be varies based on user's identity on user's identity or user data (Bordenabe 2014). The former concerns with privacy leaks in link attack to recognize user identities, such as the property in *k*-anonymity, *l*-diversity; for the latter, the detailed data is regarded as protection object, and it assumes that the adversary has known the user's identity, such as differential privacy.

Traditionally, privacy is identified based on prior experience. Now, it can be based on privacy features which are produced in different application scenarios (Xu et al. 2015).

For example, biometrics obtained from video data with highly discriminatory and uniqueness can be used as the direct conduit between the identity and data, essentially the identification of privacy (Luo et al. 2018). Moreover, with the development of machine learning, privacy can be learnt from cumulative data. These technologies can be roughly divided into two categories, one (Bordenabe 2014) is to set a trusted learner that has permission to access private database, but the way relies on the exponential mechanism. The other includes a group optimal composition chosen from the findings to user regarded as training set, and algorithms based on EM or SVM to identify privacy, even hierarchical classification.

Laws and Policies

There are three typical modes of personal information protection in the world: EU Legislative Mode, American Protection Mode, and Japanese Protection Mode. In the EU Legislative Mode, personal information is regarded as a part of individual personality and human rights. In 1995, the publication of the EU *Data Protection Directive* established the basic principles for personal information protection. In 2009, Britain issued the first personal information protection standard BS10012 *Specification for a Personal Information Management System* Young (2010). Additionally, the *General Data Protection Regulation (GDPR)* (EU) 2016/679 (European Parliament 2018b) becomes the most important law on data protection and privacy for all individuals in the European Union. American Protection Mode is based on the right to privacy. It is a combination of decentralized legislation and industry self-discipline. The *Privacy Act* (The 93th United States Congress 2018) passed in 1974 is the basic law to protect personal information privacy in the USA. Subsequently, the *Sunshine Act* (The 94th United States Congress 1976), the *E-Government Act* (The 107th United States Congress 2016), and *Consumer Privacy Bill of Right* (Meece 2012) were passed to protect personal information from different industries. Japanese Protection Mode also combines govern-

ment legislation and industry self-discipline. In 2005, in Japan the *Personal Information Protection Act* (Parliament of Canada 2003). Combined with the *Personal Information Protection Ordinance* formulated by local governments, Japan has built a relatively complete legal system for personal information protection.

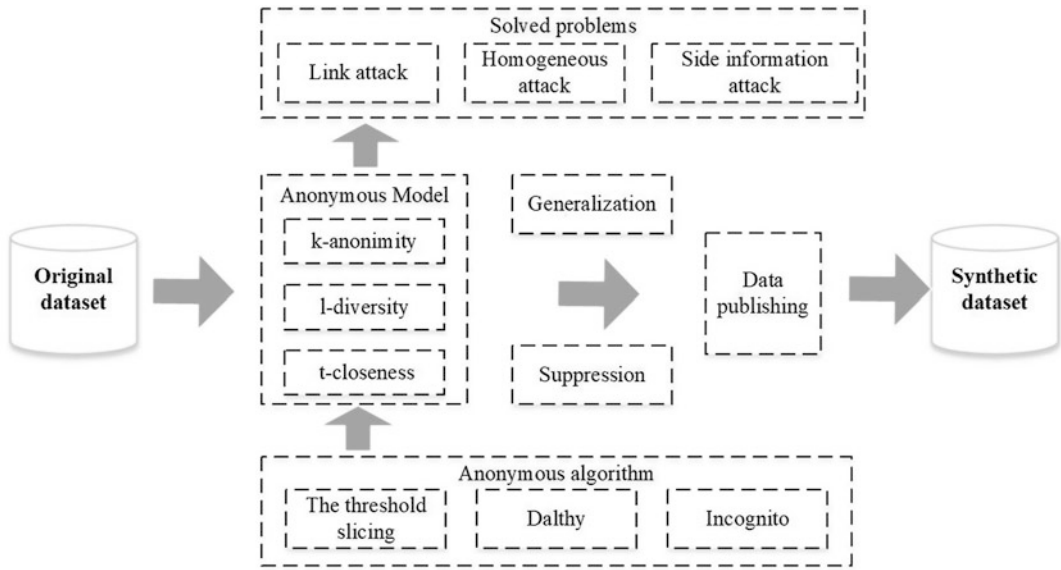
Privacy Protection Technologies

Anonymous

Regarding the privacy leakage caused by linking attack in dataset publishing, in 2002, Sweeney (2002b) proposed an anonymity model. Through the suppression and generalization (Sweeney 2002a) techniques, indistinguishability between the quasi-identity attributes is achieved. However, the existence of k -anonymity cannot resist homogeneity attack and background knowledge attack. Machanavajjhala et al. (Machanavajjhala et al. 2007) proposed l -diversity, constrained publishing data sheets, k -anonymity group sensitive attribute values to satisfy l diversification. However, the fact that l -diversity cannot resist the homogeneous attack and similarity attack, researchers have proposed an anonymous privacy protection model of t -closeness (Li et al. 2007). At present, the anonymous model has been applied in various fields including social network privacy protection, trajectory privacy protection and so on. Meanwhile, a large number of researchers extend k -anonymity to face different attacks. We will briefly summarize anonymous models and applications in Fig. 1.

Differential Privacy

Differential privacy is one of the most famous privacy notions which offer a rigorous and provable privacy guarantee for *Privacy Preserving Data Mining (PPDM)* and *Privacy Preserving Data Releasing (PPDR)* (Zhu et al. 2015). The motivation of differential privacy is to allow statistical analysis of the dataset without disclosing individual information. A formal definition is as follows (Dwork 2006, 2008; Dwork et al. 2006):



Network Privacy Surveillance and Privacy Protection, Fig. 1 Anonymous models and applications

Definition 1 A random mechanism M offers ϵ -differential privacy for any pair of D and D' , denoted as $D \sim D'$, and for every set of outcome space O , the ratio of output probabilities satisfies Eq. 1:

$$\frac{Pr[M(D) \in O]}{Pr[M(D') \in O]} \leq \exp(\epsilon) \quad (1)$$

Privacy budget ϵ determines the effect of privacy protection. The smaller the parameter ϵ is, the more noise. As such, an appropriate ϵ is the key to tradeoff accuracy and privacy. There are two basic methods to achieve differential privacy, namely *Laplace Mechanism* and *Exponential Mechanism*. But no matter which one is chosen, the global sensitivity is used to calibrate noise. For a query $Q : D \rightarrow R^k$, the *Global Sensitivity* is defined as

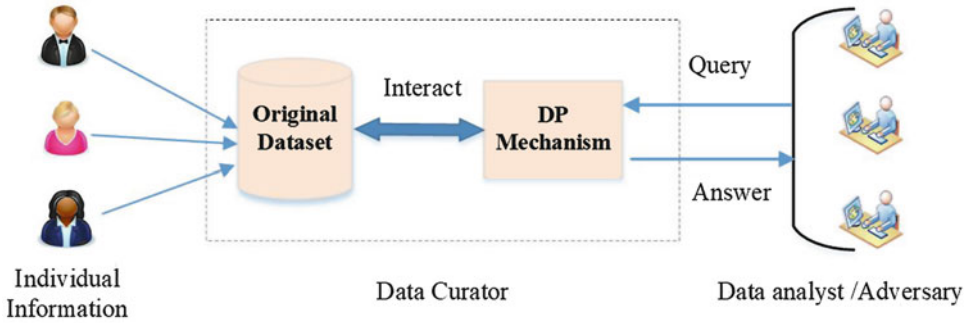
$$GS = \max_{D, D'} \|Q(D) - Q(D')\|_1 \quad (2)$$

Differential privacy can be classified into two categories: interactive and non-interactive. In a non-interactive setting the curator publishes a sanitized version of original dataset in order to protect individual information. While in the

interactive setting, trusted data curator receives queries from the analyst, adopts differential privacy mechanism and answers the queries with adaptive selection of noise distribute (e.g. Laplace distribute, Exponential distribution). Here, we use Fig. 2 to illustrate these differential privacy interactive settings.

Cryptographic Technologies

Cryptography plays an important role in privacy protection since the confidentiality and integrity that cryptography guarantees. Homomorphic encryption (Gentry 2009) maintains consistency of computation over plaintext and ciphertext. By leveraging homomorphic encryption, computation can be carried out directly over ciphertexts without any decryption. That prevents any adversary from observing the corresponding plaintext during computations. Another substitutable cryptographic tool that provides similar features is secure multiparty computation (Yao 1986), which is interactive among multiple participants. By embedding a trapdoor into a ciphertext, multiple participants collaborate to resolve a specific function $f(x_1x_2x_3, \dots, x_n)$ without leaking x_i hold by participant P_i . Searchable encryp-



Network Privacy Surveillance and Privacy Protection, Fig. 2 Differential privacy interactive model

tion (Song et al. 2000) is proposed and applied to privacy protection while both searchability and confidentiality are vital in data management, such as encrypted database, private content-based image/information retrieval, etc. In keyword searchable encryption, a trapdoor for searching is embedded into ciphertext. Hence, a specific keyword can be matched by comparing trapdoor and ciphertext. In generally searchable encryption, novel encryption/index algorithms (order preserving encryption (Agrawal et al. 2004) and comparable encryption (Furukawa 2013), etc.) are introduced to construct secure index over massive data. Based on that index, a candidate set of searching result can be located, and further refined to derive an accurate result efficiently.

Privacy Quantification and Evaluation

Assuming that a trusted data curator holds a dataset D which about n individuals, and considering a non-interactive setting data publishing problem, the curator publishes a synthetic dataset aiming to protect individual privacy. From the perspective of information theory, differential privacy system can be seen as an information-theoretic channel which inputs original dataset D and outputs a synthetic dataset $\hat{D}$ (Alvim and Andrés 2011). We model original dataset as discrete random variable X and the receiving end observable outputs as discrete random variable Y . In this model we define privacy information source entropy $H(X)$ to

evaluate the uncertainty of original dataset. For the receiving an end observable outputs Y , we introduce the conditional entropy $H(X/Y)$ to evaluate remaining uncertainty after observed outputs Y , and quantitative leaking information in the information-theoretic channel as the mutual information $I(X; Y) = H(X) - H(X/Y)$ (Peng et al. 2016).

The Balance Between Privacy Protection and Usability

In practical applications, the privacy protection level of service data is inconsistent with its availability. For example, in dummy-based location privacy protection methods (Liu et al. 2017), with the increase of privacy protection level of the user location, which means the number of dummies is increasing, the time required for the service provider to check the database is improving constantly. In l -diversity data release privacy protection methods, with the increase of l , the fine-grained of released data is decreasing, thus reducing the data's availability (Cicek et al. 2014). Therefore, it is a hot topic in current research about how to trade off the privacy protection level of service data and its availability. Shown as follows, a game model between the data provider and the data user can be used to characterize this contradictory relationship.

Definition 2 (The game model between privacy protection level of service data and its availability). The game model between privacy protection

level of service data and its availability is a three-tuple $G = \{PAU\}$, where:

- $P = \{P_{Provider}, P_{User}\}$ is the set of rational participants. $P_{Provider}$ is the data provider, and P_{User} is the data user.
- $A = \{A_{Provider}, A_{User}\}$ is the set of rational participants' strategies. $A_{Provider} = a_{Provider}(\lambda)$ is the strategy set of the data provider in which $a_{Provider}(\lambda)$ represents the strategy chosen by the data provider $P_{Provider}$ based on his privacy protection requirements. $A_{User} = a_{User}(\lambda)$ is the data user's decision on whether to use the data sent by the data provider for statistical results.
- $U = \{u_{Provider}, u_{User}\}$ is the utilities set of rational participants. Based on the data provided by the data provider $P_{Provider}$, $u_{Provider} = Sec(\lambda) + QoS(\lambda)$ is the privacy protection and service quality enjoyed by him. And $u_{User} = u(\lambda)$ represents the data user's benefits, such as the privacy information inferred of the data provider $P_{Provider}$.

Key Applications

At present, the opening and sharing of data lead to a large number of privacy issues. Thus, some privacy-preserving applications works are required to solve them, such as anonymous processing, differential privacy processing, etc. In addition, to resolve the privacy disclosure issues in social networks, in some applications intelligent recommendations are included for privacy preference settings to their users. Last but not least, in many applications risk assessment of privacy leakage is also included before the users make decision on their on-line use.

Future Directions

In the future, with the development of artificial intelligence (AI), the traditional privacy preserving method may no longer be suitable for the application demands. To solve the problem of

the personal privacy can be sensed by machine learning technologies, and the methods based on generative adversarial networks will be the future directions of personal privacy preserving. The generative adversarial networks may be able to interfere machine learning models by adding a tiny and imperceptible perturbation to lead them misclassifying. Furthermore, to resist the risk of privacy leakage caused by AI technologies, some legislation and supervision cannot be avoided.

Cross-References

► Differential Privacy

Acknowledgments The work is supported by the National Natural Science Foundation of China under grant No.: 61662009, and the National Cryptography Development Fund under grant No. MMJJ20170129.

References

- Agrawal R, Kiernan J, Srikant R, Xu Y (2004) Order preserving encryption for numeric data. In: Proceedings of the 2004 ACM SIGMOD international conference on management of data, SIGMOD'04. ACM, New York, pp 563–574. <https://doi.org/10.1145/1007568.1007632>
- Alvim MS, Andrés ME (2011) On the relation between differential privacy and quantitative information flow. In: Proceedings of the 38th international conference on automata, languages and programming – volume part II, ICALP'11. Springer, Berlin/Heidelberg, pp 60–76. <http://dl.acm.org/citation.cfm?id=2027223.2027228>
- American Civil Liberties Union (2013) Privacy & technology. <https://www.aclu.org/issues/privacy-technology>
- Ashrafi N, Kuilboer JP (2018) A comparative study of privacy protection practices in the US, Europe, and Asia. Int J Inf Secur Priv (IJISP) 12(3):1–15. <https://doi.org/10.4018/IJISP.2018070101>
- Bordenabe NE (2014) Measuring privacy with distinguishability metrics: definitions, mechanisms and application to location privacy. Phd thesis
- Cicek AE, Nergiz ME, Saygin Y (2014) Ensuring location diversity in privacy-preserving spatio-temporal data publishing. VLDB J 23(4):609–625. <https://doi.org/10.1007/s00778-013-0342-x>
- Dwork C (2006) Differential privacy. In: Proceedings of the 33rd international conference on automata, languages and programming – volume part II, ICALP'06. Springer, Berlin/Heidelberg, pp 1–12. https://doi.org/10.1007/11787006_1
- Dwork C (2008) Differential privacy: a survey of results. In: Proceedings of the 5th international conference

- on theory and applications of models of computation, TAMC'08. Springer, Berlin/Heidelberg, pp 1–19. <http://dl.acm.org/citation.cfm?id=1791834.1791836>
- Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Proceedings of the third conference on theory of cryptography, TCC'06. Springer, Berlin/Heidelberg, pp 265–284. https://doi.org/10.1007/11681878_14
- European Parliament, Council of the European Union (2018b) General Data Protection Regulation. [https://en.wikipedia.org/w/index.php?title=\\$General_Data_Protection_Regulation&oldid=847469216](https://en.wikipedia.org/w/index.php?title=$General_Data_Protection_Regulation&oldid=847469216), page Version ID: 847469216
- Furukawa J (2013) Request-based comparable encryption. In: Crampton J, Jajodia S, Mayes K (eds) Computer security – ESORICS 2013. Springer, Berlin/Heidelberg, pp 129–146
- Gentry C (2009) Fully homomorphic encryption using ideal lattices. In: Proceedings of the forty-first annual ACM symposium on theory of computing, STOC'09. ACM, New York, pp 169–178. <https://doi.org/10.1145/1536414.1536440>
- Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd international conference on data engineering, pp 106–115. <https://doi.org/10.1109/ICDE.2007.367856>
- Liu H, Li X, Li H, Ma J, Ma X (2017) Spatiotemporal correlation-aware dummy-based privacy protection scheme for location-based services. In: IEEE INFOCOM 2017 – IEEE conference on computer communications, pp 1–9. <https://doi.org/10.1109/INFOCOM.2017.8056978>
- Luo Y, Cheung SCS, Lazeretti R, Pignata T, Barni M (2018) Anonymous subject identification and privacy information management in video surveillance. *Int J Inf Secur* 17(3):261–278. <https://doi.org/10.1007/s10207-017-0380-2>
- Lyon D, Zureik E (eds) (1996) Computers, surveillance, and privacy, 1st edn. University Of Minnesota Press, Minneapolis
- Machanavajjhala A, Kifer D, Gehrke J, Venkatasubramanian M (2007) L-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Discov Data* 1(1). <https://doi.org/10.1145/1217299.1217302>
- Meece M (2012) President Obama's consumer privacy bill of rights. <https://www.forbes.com/sites/mickeymeece/2012/02/23/president-obamas-consumer-privacy-bill-of-rights/>
- Parliament of Canada (2003) Personal Information Protection Act. http://www.bclaws.ca/Recon/document/ID/freeside/00_03063_01
- Peng C, Ding H, Zhu Y, Tian Y, Fu Z (2016) Information entropy models and privacy metrics methods for privacy protection. *J Softw* 27(8): 1891–1903 (in Chinese). <https://doi.org/10.13328/j.cnki.jos.005096>, <http://www.jos.org.cn/1000-9825/5096.htm>
- Song DX, Wagner D, Perrig A (2000) Practical techniques for searches on encrypted data. In: Proceedings of the 2000 IEEE symposium on security and privacy, SP'00. IEEE Computer Society, Washington, pp 44–45. <http://dl.acm.org/citation.cfm?id=882494.884426>
- Sweeney L (2002a) Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertain Fuzziness Knowl Based Syst* 10(5):571–588. <https://doi.org/10.1142/S021848850200165X>
- Sweeney L (2002b) K-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst* 10(5):557–570. <https://doi.org/10.1142/S0218488502001648>
- The 93th United States Congress (2018) Family Educational Rights and Privacy Act. https://en.wikipedia.org/w/index.php?title=Family_Educational_Rights_and_Privacy_Act&oldid=845699239, page Version ID: 845699239
- The 94th United States Congress (1976) Government in the Sunshine Act. <https://www.gsa.gov/policy-regulations/policy/federal-advisory-committee-management/legislation-and-regulations/government-in-the-sunshine-act>
- The 107th United States Congress (2016) E-Government Act of 2002. <https://www.archives.gov/about/laws/egov-act-section-207.html>
- Warren SD, Brandeis LD (1890) The right to privacy. *Harvard law review*, pp 193–220
- Westin AF (2003) Social and political dimensions of privacy. *J Soc Issues* 59(2):431–453. <https://doi.org/10.1111/1540-4560.00072>
- Xu K, Yue H, Guo L, Guo Y, Fang Y (2015) Privacy-preserving machine learning algorithms for big data systems. In: 2015 IEEE 35th international conference on distributed computing systems, pp 318–327. <https://doi.org/10.1109/ICDCS.2015.40>
- Yao ACC (1986) How to generate and exchange secrets. In: Proceedings of the 27th annual symposium on foundations of computer science, SFCS'86. IEEE Computer Society, Washington, pp 162–167. <https://doi.org/10.1109/SFCS.1986.25>
- Young L (2010) BS 10012:2009 Data protection specification for a personal information management system. *Rec Manage J* 20(1). <https://doi.org/10.1108/rmj.2010.28120aee.003>
- Zhu T, Xiong P, Li G, Zhou W (2015) Correlated differential privacy: hiding information in non-iid data set. *IEEE Trans Inf Forensics Secur* 10(2):229–242. <https://doi.org/10.1109/TIFS.2014.2368363>

Network Security

- [Blockchain: Enabling Future Internet with Intrinsic Security](#)

Network Selection in Heterogeneous Wireless Networks

Wei Song

University of New Brunswick, Fredericton, NB, Canada

Synonyms

Call assignment; Radio access selection

Definition

When the cellular networks are interconnected with the wireless local area networks (WLANs), there is a two-tier overlaying structure which offers both cellular access and WLAN access to a dual-mode mobile device within the WLAN-covered area. Ubiquitous coverage is provided by the cellular network (higher-tier), while WLANs (lower-tier) are deployed in disjoint hot-spot areas. Then, there comes the network selection problem of how to properly admit incoming traffic to the cell or WLAN. Specifically, a preferred target network, either a cellular cell or a WLAN, should be first selected based on various decision criteria by taking into account factors such as service type and network conditions of the two networks. A service request rejected by its first-choice network can just leave the system or further try to access the other network.

Historical Background

Nowadays, the wireless networks support heterogeneous wireless access technologies. The popular cellular networks and WLANs present perfectly complementary characteristics in terms of service capacity, mobility support, and quality-of-service (QoS) provisioning. On one hand, the fourth-generation (4G) cellular networks are spreading fast, providing ubiquitous wide-area coverage, QoS-guaranteed services, and

full-spectrum mobility support. On the other hand, IEEE 802.11-based WLANs are widely deployed in hotspot areas (e.g., offices, airports, and hotels), supporting high-rate and low-cost data services in the restricted small areas. The cellular/WLAN interworking offers an effective approach for sharing load and saving cost.

Many integration architectures have been proposed for cellular/WLAN interworking. According to the inter-dependence between the two networks, the integration architectures are classified into two categories, i.e., tight coupling and loose coupling. The interworking can be very tight with an integration at the radio access network (RAN) level or less tight with the integration in the cellular core network (CN). The interworking can also be loose with the two networks integrated beyond the core network and usually through an external Internet protocol (IP) network.

With tight coupling, the existing cellular infrastructure can be reused to a large extent, and it is possible to apply joint resource allocation to optimize overall system capacity. The main disadvantage of tight coupling is the high implementation complexity. An interworking unit equipped with a cellular-compatible interface is necessary to expose the WLAN to the cellular core or even the cellular RAN. On the other hand, the loose-coupling architecture can employ the pervasive IP technology to glue together the cellular network and WLANs. Although imposing minimal modification to the WLAN, loose coupling requires the cellular network to be augmented with extra functionalities such as Mobile IP. Nonetheless, loose coupling is relatively inefficient because of a long signaling path, redundant processing in the two networks, and a large number of network elements involved in management operations.

Foundations

With the cellular/WLAN interworking, many challenging research issues arise to take advantage of the complementary characteristics of the two different networks so as to enhance their strength and compensate for their limitation. The

key research problems include vertical handoff (i.e., the handoff between wireless networks of different access technologies), network selection, call admission control, and resource allocation. In fact, these research problems are closely related and need to be addressed jointly to achieve maximum interworking gain.

While the cellular network provides ubiquitous connectivity with wide-area coverage, WLANs are only deployed disjointly in hotspot areas. The cellular/WLAN interworking then results in an overlay structure, which offers both cellular access and WLAN access to dual-mode mobiles in the overlay area. First, when there is an incoming request, either an available cell or WLAN should be selected to accommodate the request. However, the selected network may accept or reject the call depending on its available resources and the admission control policy in place. For an unsuccessful initial assignment, the call can resort to the other network for access or just quit. In addition, existing calls in the integrated system can be dynamically migrated between the cellular part and the WLANs. For example, when sufficient resources are released from call completion or outgoing handoff, an existing call can be reallocated to a more desirable network according to current network conditions. The call reassignment is also referred to as take-back in some literature.

Together with the aforementioned network selection strategies, the admission control policies in the target networks need to be properly designed to limit the admissible traffic load and provide QoS assurance. It is vital to address network selection and admission control jointly in order to optimize the traffic placement and maximize the overall resource utilization. For example, as handoff dropping is more annoying than new call blocking, handoff calls should be prioritized over new calls in the admission control policy, e.g., by means of reserving guard channels, queueing handoff calls, and so on. Correspondingly, it is preferable to distinguish handoff calls from new calls in the network selection strategy as well. In addition, user location is another key factor for both network selection and admission control, since the accessible resources vary with locations.

Specifically, in order to efficiently utilize the integrated resources, the following unique characteristics of the integrated network should be considered in network selection and admission control.

- **Heterogeneous wireless access environment:** Aimed at different applications, the cellular networks differ from WLANs at the physical layer, medium access, and link control layers. For example, the cellular networks use allocation-based channel access such as code-division multiple access (CDMA) and orthogonal frequency-division multiple access (OFDMA). The centralized control and reservation-based resource allocation at base stations (BSs), together with proper admission control to limit the traffic load, enable fine QoS provisioning in cellular networks. In contrast, the channel access of most popular WLANs is contention-based random access, e.g., the mandatory distributed coordination function (DCF) of IEEE 802.11 WLAN. DCF adopts carrier sense multiple access with collision avoidance (CSMA/CA) and binary exponential backoff. The access point also needs to compete for access with other mobile devices and therefore behaves differently from the BS. Correspondingly, the service provisioning is in a best-effort manner without QoS assurance.
- **Hierarchical overlay network structure:** After four generations of evolution, the cellular networks have widely entrenched infrastructure, which provides almost ubiquitous connectivity and supports full-spectrum user mobility from fast highway vehicles to stationary users in an indoor environment. In contrast, WLANs are usually deployed disjointly in hotspot local areas, where the traffic intensity is typically much higher than surrounding areas. Thus, the cellular/WLAN interworking results in an overlay structure. Both cellular access and WLAN access are available to mobiles within WLAN-covered areas. The incoming traffic load should be properly shared between the overlay cell and WLAN for QoS enhancement, congestion relief, cost reduction, etc.

- **Multi-service traffic load:** Multi-service support has become an essential requirement for current wireless networks. Different services usually require different QoS deliveries. Real-time service such as voice and video are sensitive to delay, while the main concern for delay-tolerant data service is the throughput. Multiple services can take a good advantage of the complementary strength of the two networks. For example, benefiting from the centralized infrastructure, cellular networks can effectively serve real-time traffic with stringent QoS requirements. On the other hand, WLANs can be a good choice for elastic data service, which may experience traffic asymmetry at the uplink and downlink. The contention-based access of WLANs enables a virtually time division duplexing (TDD) mode to efficiently handle the load asymmetry and flexibly adapt to traffic elasticity. As seen, the service type is an important factor in network selection for cellular/WLAN interworking.
- **Location-dependent user mobility:** User mobility has been extensively studied for resource allocation in hierarchical cellular networks where microcells are overlaid with macrocells. However, many design principles are not applicable to cellular/WLAN interworking, although there exists a similar overlay structure. For WLANs deployed in indoor hotspots such as offices and hotels, the low user mobility within these areas results in a heavy-tailed residence time (Thajchayapong and Peha 2006). Consequently, a uniform mobility model becomes invalid for a large cell. As the user residence time in a cell or WLAN directly affects channel holding time, this new characteristic further complicates network selection.

Next, we introduce some representative approaches for network selection with cellular/WLAN interworking. These solutions often address network selection together with related issues such as vertical handoff, connection-level bandwidth allocation, call admission control, and load sharing.

Song et al. analyzed a simple and easy-to-implement network selection strategy, referred

to as WLAN-first scheme, where WLANs are always preferred by all services whenever the WLAN access is available, so as to take advantage of the low cost and large bandwidth of WLANs (Song et al. 2007b). An incoming service request rejected by a WLAN overflows to the cellular network to request admission if it is a new call, or remains in the cellular network if it is an ongoing call carried by the overlaying cell. Considering real-time voice service and interactive data service (such as Web browsing), this work derives the admission regions for the integrated network.

A service-differentiated network selection strategy is investigated in Song et al. (2005). Considering the distinct features of the cellular network and WLANs, it is essential to apply different network selection policies for delay-sensitive real-time traffic and delay-tolerant elastic data traffic. Due to the large cell size and ubiquitous coverage of the cellular network, it is preferable to first assign real-time voice calls to the cellular network. On the other hand, interactive data calls that accept elastic bandwidth can more efficiently utilize the large WLAN bandwidth. Hence, data service should attempt to access the WLAN first and only resort to the cellular network when the WLAN is congested.

To further reduce implementation complexity and enable flexible control, the service-differentiated principle in network selection can also be incorporated in a randomized selection policy. As proposed in Song et al. (2007a), an incoming voice (resp. data) call in the overlay area can select the covering cell or the WLAN according to a probability θ_v^c (resp. θ_d^c) and θ_v^w (resp. θ_d^w), respectively. Here, it is obvious that $\theta_v^c = 1 - \theta_v^w$ and $\theta_d^c = 1 - \theta_d^w$. The key problem is to determine these selection parameters according to traffic and mobility estimates. In addition, these parameters can be adapted to control the load sharing between the two networks. For example, θ_d^w can be increased appropriately if more data traffic needs to be offloaded to the WLANs.

In addition to the service type, many other factors can be considered in the network selection problem, e.g., user moving speed, chan-

nel condition, traffic load status, mobile device type, personal preference, etc. Correspondingly, different techniques can be applied to assimilate these factors and rank the candidate access options. The decision process in Bari and Leung (2007) uses noncompensatory and compensatory multiattribute decision-making (MADM) algorithms jointly on the network side to assist the mobile device to select the top candidate network. There are different MADM solutions such as multiplicative exponent weighting (Yoon and Hwang 1995), simple additive weighting (Zhang 2004), and the technique for order preference by similarity to ideal solution (TOPSIS) (Zhang 2004).

The network selection scheme in Song and Jamalipour (2005) is based on an integrated algorithm of analytic hierarchy process (AHP) and gray relational analysis (GRA). AHP quantitatively weights decision alternatives by hierarchical and pairwise comparison, while GRA ranks network alternatives efficiently through building a gray relationship with an ideal option.

Gelabert et al. studied radio access selection in heterogeneous multiaccess multiservice wireless networks (Gelabert et al. 2008). A 4D Markov chain is proposed accounting for two service types, generically voice and data, which are served over two access networks. The proposed analytical framework can be used to define a wide range of access selection policies taking into account several criteria such as service type, network conditions, and device type. Two service-based network selection schemes are compared with a load balancing scheme, a terminal-driven scheme, and a random policy. Especially, the service-based schemes indicate a tradeoff between the average data throughput per user and the total blocking probability.

In Si et al. (2010), application-layer multimedia QoS (e.g., video distortion) is considered as decision criteria for network selection, which is different from the focus of network-layer QoS, such as blocking probability, throughput, and resource utilization efficiency. The authors formulate the heterogeneous wireless networks as a restless bandit system, which has been widely

applied in operations research and stochastic control. With the restless bandit approach, the optimal network selection policy is derived in a fully distributed and scalable manner. Moreover, the application layer parameter, intra-refreshing rate, is adapted to further improve user experience with multimedia applications.

Game theory is another popular technique to investigate network selection with connection-level bandwidth allocation and admission control for heterogeneous wireless networks. Niyato and Hossain present a game-theoretic framework for a heterogeneous wireless access environment, in which a mobile user is able to connect to a WLAN, a wireless metropolitan area network (WMAN), and a cellular network simultaneously (Niyato and Hossain 2008). Two noncooperative games are formulated for the network-level bandwidth allocation to different service areas, and the connection-level admission control to arriving connections, respectively. Further, a bargaining game is formulated to obtain the capacity reservation thresholds so that vertical and horizontal handoff connections are prioritized over new connections.

In Yu and Krishnamurthy (2007), an optimal joint session admission control scheme is proposed for multimedia traffic in the cellular/WLAN integrated network. It is based on a semi-Markov decision process (SMDP) to maximize the overall network revenue with QoS constraints. Throughput and packet delay are considered as QoS constraints in the WLAN, while signal-to-interference (SIR) and SIR outage probability are considered for the CDMA based cellular network. An infinite horizon SMDP is formulated and solved by the linear-programming-based algorithms. The optimal joint admission control policy obtained thereby is found to be a randomized policy.

Key Applications

Network selection for heterogeneous wireless networks is fundamental in network resource allocation, QoS assurance, load sharing, vertical handoff management, and call admission control

so as to optimize multi-service provisioning and maximize overall resource utilization efficiency.

Cross-References

- [Matching Game for Load Distribution in Multi-tier Cellular Networks](#)
- [Mobile Edge Caching in HetNets](#)
- [WiFi Offloading in WiFi-Cellular Networks](#)

References

- Bari F, Leung VCM (2007) Automated network selection in a heterogeneous wireless network environment. *IEEE Netw* 21(1):34–40
- Gelabert X, Pérez-Romero J, Sallent O, Agustí R (2008) A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks. *IEEE Trans Mob Comput* 7(10):1257–1270
- Niyato D, Hossain E (2008) A noncooperative game-theoretic framework for radio resource management in 4G heterogeneous wireless access networks. *IEEE Trans Mob Comput* 7(3):332–345
- Si P, Ji H, Yu FR (2010) Optimal network selection in heterogeneous wireless multimedia networks. *ACM/Springer Wirel Netw* 16(5):1277–1288
- Song Q, Jamalipour A (2005) Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques. *IEEE Wirel Commun* 12(3):42–48
- Song W, Jiang H, Zhuang W, Shen X (2005) Resource management for QoS support in cellular/WLAN interworking. *IEEE Netw* 19(5):12–18
- Song W, Cheng Y, Zhuang W (2007a) Improving voice and data services in cellular/WLAN integrated network by admission control. *IEEE Trans Wirel Commun* 6(11):4025–4037
- Song W, Jiang H, Zhuang W (2007b) Performance analysis of the WLAN-first scheme in cellular/WLAN interworking. *IEEE Trans Wirel Commun* 6(5):1932–1952
- Thajchayapong S, Peha J (2006) Mobility patterns in microcellular wireless networks. *IEEE Trans Mob Comput* 5(1):52–63
- Yoon K, Hwang C (1995) Multiple attribute decision making: an introduction. Sage, Thousand Oaks
- Yu F, Krishnamurthy V (2007) Optimal joint session admission control in integrated WLAN and CDMA cellular network. *IEEE Trans Mob Comput* 6(1):126–139
- Zhang W (2004) Handover decision using fuzzy MADM in heterogeneous networks. In: *Proceedings of the IEEE Wireless Communications and Networking (WCNC)*, vol 2, pp 653–658

Network Sharing

- [Sharing Economics](#)

Network Slicing-Enabled Green C-RAN

Yi-Han Chiang and Yusheng Ji
Information Systems Architecture Science
Research Division, National Institute of
Informatics, Tokyo, Japan

Synonyms

[Green C-RAN with network slicing](#)

Definition

Network slicing is an emerging technology that allows mobile network operators (MNOs) to slice their physical networks into logical ones, thereby providing independent services to multiple mobile virtual network operators (MVNOs). With the incorporation of network slicing, physical network resources can be allocated in terms of network slices, each of which represents a virtualized and isolated network with reserved network resources.

Cloud radio access network (C-RAN) incorporates the concept of cloud computing into RANs, which generally consists of a baseband unit (BBU) pool and remote radio heads (RRHs). The BBU pool is located at a centralized site with abundant computing resources, while the RRHs are distributed antenna units that are deployed apart from the BBU pool. By pooling the BBUs of RRHs together, lower capital and operation expenditures (CapEx and OpEx) for MNOs can be envisaged.

Remote radio head (RRH) is a set of lightweight antenna units with its processing units partly or totally migrated to the BBU pool, but retaining its location at distant sites. RRHs are

connected to the BBU pool by delivering in-phase and quadrature (IQ) data through fronthaul (FH) links. Thanks to the resource pooling feature, the use of RRHs evolves as a cost-effective alternative (due to the simplified hardware architectures as compared to traditional macro base stations) and makes large-scale network deployments easier than before.

Historical Background

The emergence of mobile data tsunami as forecasted in Cisco (2019) has stimulated MNOs to search for next-generation wireless networks. To catch up with the explosive growth of mobile data traffic while keeping CapEx and OpEx at a minimum, network densification has been widely regarded as a promising solution in that it offers denser service coverage and higher capacity due to the massively deployed small cells, and benefits from lower transmit power consumption because of shortened last-mile distances. To manage such denser networks efficiently, MNOs are inspired to pay attention to C-RAN (I et al. 2014; Checko et al. 2015; Saxena et al. 2016; Alimi et al. 2018) in its benefits of better resource optimization and cost effectiveness.

In parallel with the standardization of 5G systems, network slicing (Zhou et al. 2016; Rost et al. 2017; Foukas et al. 2017; Zhang et al. 2017) has attracted growing research attention due to its flexibility and scalability to meet diverse requirements of future mobile networks. By means of network slicing, each physical network (owned by an MNO) can be dynamically sliced into several logical ones, thereby providing reserved network resources to individual MVNOs. In this way, a wide range of services (e.g., data, video and telephony) with distinct requirements (e.g., bandwidth, delay, reliability and privacy) from different MVNOs can be embedded in the same physical network.

In addition to boosting network capacity, plenty of existing works have been devoted to energy-saving issues (Arnold et al. 2010; Auer et al. 2011) in cellular systems, thereby saving CapEx and OpEx for MNOs. Thanks to the

centralized architecture, C-RAN can be smoothly integrated with network slicing while achieving network greenness. Since BBUs and RRHs will be initiated to serve user equipment (UE) for MVNOs, where BBUs and RRHs can incur power consumption in their computation and communication, respectively, they should be efficiently activated and utilized. One possibility for the energy conservation inspired by Wang et al. (2017), Aqeeli et al. (2018) and Yao and Ansari (2018) is that one BBU can serve a cluster of RRHs, and less active BBUs can lead to lower power consumption in the BBU pool. Another energy-saving opportunity from Ashraf et al. (2011), Holtkamp et al. (2014) and Chiang and Liao (2017) indicates that each RRH consumes power in its operation and radio transmission. As a consequence, the less the active RRHs, the lower the resulting power consumption. Altogether, how BBUs and RRHs can be activated and how BBU-RRH and RRH-UE mappings can be performed, while satisfying network capacity limits and user requirements, are decisive in green C-RAN with network slicing.

Foundations

A generic C-RAN architecture (as shown in Fig. 1) consists of a set $\mathcal{L}$ of BBUs (colocated in a BBU pool) and a set $\mathcal{M}$ of RRHs, where each RRH is connected to the BBU pool through a high-speed FH link, and the BBUs are controlled and managed by the MNO. On top of the C-RAN, there exists a set $\mathcal{V}$ of MVNOs, each of which has an independent set $\mathcal{N}_v$ of UE (each with a data rate requirement of R_n , $\forall n \in \mathcal{N}_v$) to serve. After receiving service requests from MVNOs, the MNO needs to serve the set $\mathcal{N} = \bigcup_{v \in \mathcal{V}} \mathcal{N}_v$ of UE.

In the BBU pool, an active BBU can be realized by initiating a virtual machine (VM), which can serve one or more UE. Clearly, the BBU-RRH mapping is either one-to-one or one-to-many. In addition, each BBU $l \in \mathcal{L}$ can only support up to $C_l^{\max}$ RRHs, due to its limited computing resources and FH capacities.

Each BBU/RRH can be activated or deactivated to reduce its operational power

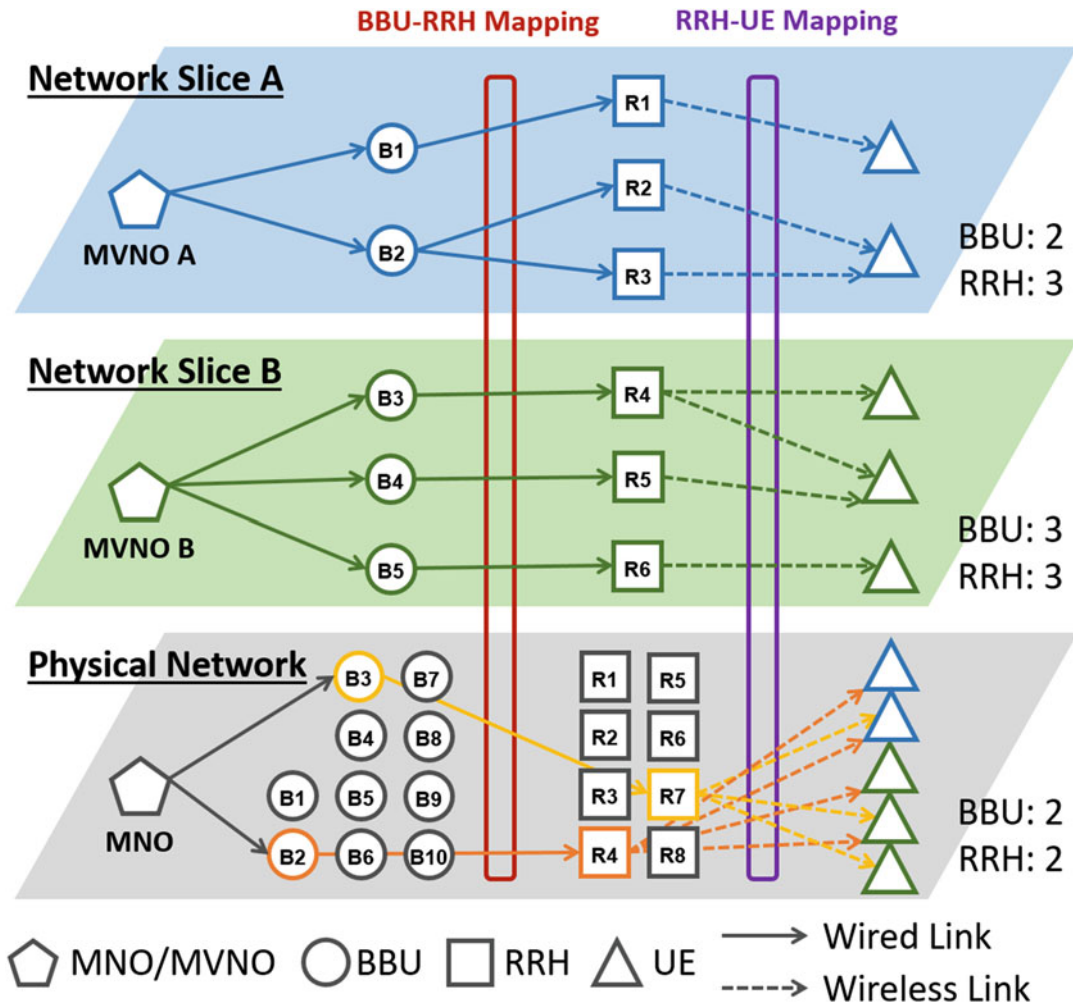
consumption. Typically, such a deactivation can save energy effectively, since a non-zero energy dissipation can be attributed to operating in active mode. Therefore, operational power consumption can be lowered whenever a BBU/RRH can be deactivated.

Thanks to the centralized processing of C-RAN, multiple RRHs can cooperatively transmit to UE, thereby improving the received signal qualities. Suppose that universal frequency reuse is adopted by all RRHs to fully utilize the network spectrum of bandwidth B . Denote by SINR_n the received signal-to-interference-plus-

noise ratio (SINR) at each UE as

$$\text{SINR}_n = \frac{\sum_{m \in \mathcal{M}} p_{mn} H_{mn}}{I_n + N_0}, \quad \forall n \in \mathcal{N}, \quad (1)$$

where H_{mn} refers to the channel power gain from RRH m to UE n , N_0 is the noise spectral density at each UE, I_n aggregates the interfering powers experienced by UE n . In addition, $p_{mn} \leq P_m^{\text{tx}}$ is the transmit power allocated to UE n by RRH m , where P_m^{tx} is the RF output power of RRH m at its maximum load (namely RRH m 's transmit power budget).



Network Slicing-Enabled Green C-RAN, Fig. 1 An illustration of a network slicing-enabled C-RAN, where MVNO A requests for 2 BBUs and 3 RRHs, and MVNO B asks for 3 BBUs and 3 RRHs. After collecting

the demands from the MVNOs, the MNO simply activates 2 BBUs and 2 RRHs, in the way that network capacity limits and user requirements can all be met

The total power consumption is to aggregate the operational and transmit power consumption

of all BBUs and RRHs, which can be expressed as

$$\begin{aligned}
 p^{\text{tot}} &= \overbrace{\sum_{l \in \mathcal{L}} P_{R,m}^{\text{act}} y_m + \sum_{l \in \mathcal{L}} P_{R,m}^{\text{slp}} (1 - y_m) + \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \Delta_{R,m} p_{mn}}^{\text{RRH power consumption}} \\
 &+ \underbrace{\sum_{m \in \mathcal{M}} P_{B,l}^{\text{act}} z_l + \sum_{m \in \mathcal{M}} P_{B,l}^{\text{slp}} (1 - z_l) + \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}} \Delta_{B,l} q_{lm}}_{\text{BBU power consumption}},
 \end{aligned} \tag{2}$$

where $P_{R,m}^{\text{act}}$ and $P_{R,m}^{\text{slp}}$ are the power consumption of RRH m operating in active mode and sleep mode, respectively, and $P_{B,l}^{\text{act}}$ and $P_{B,l}^{\text{slp}}$ are those of BBUs. $\Delta_{R,m}$ and $\Delta_{B,l}$ represent the slopes of the load-dependent power consumption of RRH m and BBU l , respectively. The binary decision variables y_m and z_l indicate the activeness of RRH m and BBU l , p_{mn} is the transmit power of RRH m allocated for serving UE n , and q_{lm} indicates whether RRH m is associated with BBU l .

Mathematically, the problem of enabling network slicing to green C-RAN (NSGC) can be formulated as (for brevity, $\mathbf{p} = \{p_{mn}, \forall m, n\}$, $\mathbf{q} = \{q_{lm}, \forall l, m\}$, $\mathbf{y} = \{y_m, \forall m\}$ and $\mathbf{z} = \{z_l, \forall l\}$):

$$\begin{aligned}
 \min_{\mathbf{p}, \mathbf{q}, \mathbf{y}, \mathbf{z}} \quad & p^{\text{tot}}, \\
 \text{s.t.} \quad & \mathbf{C1: } B \log(1 + \text{SINR}_n) \geq R_n, \forall n \in \mathcal{N}, \\
 & \mathbf{C2: } \sum_{n \in \mathcal{N}} p_{mn} \leq P_m^{\text{tx}} y_m, \forall m \in \mathcal{M}, \\
 & \mathbf{C3: } \sum_{l \in \mathcal{L}} q_{lm} = y_m, \forall m \in \mathcal{M}, \\
 & \mathbf{C4: } \sum_{m \in \mathcal{M}} q_{lm} \leq C_l^{\text{max}} z_l, \forall l \in \mathcal{L}, \\
 & \mathbf{C5: } p_{mn} \geq 0, \forall m \in \mathcal{M}, n \in \mathcal{N}, \\
 & \mathbf{C6: } q_{lm}, y_m, z_l \in \{0, 1\}, \forall l \in \mathcal{L}, m \in \mathcal{M},
 \end{aligned}$$

where **C1** states that the achieved data rate should be high enough to meet each UE's data rate requirement, **C2** indicates that each RRH can allocate transmit powers to UE (up to its transmit power budget) only if it is operating in active mode, **C3** ensures that each active RRH is associated with exactly one BBU, **C4** presents that each active BBU can only accommodate a limited amount of RRHs, **C5** and **C6** refer to the decision variables of RRH-UE mapping, BBU-RRH mapping, the activeness of RRHs and BBUs, respectively.

In fact, finding an optimal solution to the NSGC problem is essentially NP-hard. Therefore, it is of more practical interests to look for an approximate solution with a polynomial-time complexity and a provably good performance guarantee.

Definition 1 Consider the NSGC problem as follows: does there exist a solution $(\mathbf{p}, \mathbf{q}, \mathbf{y}, \mathbf{z})$ such that the constraints **C1** to **C6** can all be satisfied, and the total power consumption p^{tot} is at most U ? The problem instance can be expressed as

$$\begin{aligned}
 & \text{NSGC}(\mathcal{L}, \mathcal{M}, \mathcal{N}, B, \{R_n\}, \{H_{mn}\}, \{C_l^{\text{max}}\}, \{P_m^{\text{tx}}\}, N_0, \\
 & U, \{\Delta_{R,m}\}, \{\Delta_{B,l}\}, \{P_{R,m}^{\text{act}}\}, \{P_{B,l}^{\text{act}}\}, \{P_{R,m}^{\text{slp}}\}, \{P_{B,l}^{\text{slp}}\}).
 \end{aligned}$$

Definition 2 (WEIGHTED SET COVER) Given a collection $\mathcal{C}$ of subsets of a finite set $\mathcal{S}$ with $\bigcup_{D \in \mathcal{C}} D = \mathcal{S}$ and a positive integer V . Does $\mathcal{C}$ contain a subcollection $\mathcal{C}' \subseteq \mathcal{C}$ with $\sum_{D \in \mathcal{C}'} W_D \leq V$, such that every element of $\mathcal{S}$ belongs to at least one member of $\mathcal{C}'$? The problem instance can be denoted by $\text{WSC}(\mathcal{S}, \mathcal{C}, V, \{W_D\})$.

Theorem 1 *The NSGC problem is NP-hard.*

Proof. Consider the restricted case of only one BBU, i.e., $\mathcal{L} = \{l^\circ\}$. Since the NSGC problem involves power allocation whereas the WSC problem does not have the analogous part, it is necessary make the data rate requirements low enough to ensure that **C1** can always be held for all UE. By carefully examining **C1**, it can be observed that the following fact holds: as $R_n \geq \bar{R}_n$, **C1** can be satisfied as long as each UE is served by an arbitrary one RRH with a uniform transmit power (i.e., $P_m^{\text{tx}}/|\mathcal{N}_{l^\circ}|$ for each $n \in \mathcal{N}_{l^\circ}$), where

$$\bar{R}_n = B \log \left[1 + \left(\frac{\sum_{l \in \mathcal{L}} |\mathcal{N}_{l^\circ}| (N_0 + \sum_{m \in \mathcal{M}} P_m^{\text{tx}} H_{mn})}{\min_{m \in \mathcal{M}} P_m^{\text{tx}} \cdot \sum_{m \in \mathcal{M}} H_{mn}} - 1 \right)^{-1} \right]. \quad (3)$$

Given with $R_n \geq \bar{R}_n$, it is apparent that the uniform transmit powers are always allocable.

The NP-hardness of the NSGC problem can then be proved by restriction. In particular, the NSGC problem can be restricted by allowing only instances with simply the active operational power consumption (by setting $\Delta_{R,m}, \Delta_{B,l}, P_{R,m}^{\text{slp}}$ and $P_{B,l}^{\text{slp}}$ to zeros) and binary channel states ($H_{mn} \in \{0, 1\}$). Other parameters can be chosen freely without affecting the following instance mapping. Now, consider an WSC problem instance of

$$\text{WSC}(\mathcal{S}, \mathcal{C}, V, \{W_D\}).$$

Then, the corresponding restricted NSGC problem instance can be obtained as

$$\text{NSGC} \left(\{l^\circ\}, \mathcal{C}, \mathcal{S}, 1, \{K_1, K_1, \dots\}, \{E_{Cs}\}, K_2, \{1, 1, \dots\}, |\mathcal{C}|, V, \{0, 0, \dots\}, \{0, 0, \dots\}, \{W_D\}, \{W_D\}, \{0, 0, \dots\}, \{0, 0, \dots\} \right).$$

where $K_1 = \log(|\mathcal{S}|/|\mathcal{S}|-1)$, $K_2 = |\mathcal{C}|$, $E_{Ds} = 1$ if $s \in D$ and 0 otherwise. Since the transformation between the restricted NSGC problem instance and the WSC problem instance can be run in polynomial time, the NP-hardness of the NSGC problem is thus proved. $\square$

Key Applications

Network slicing-enabled green C-RAN generally involves the dynamic operations of BBU/RRH activation, BBU-RRH and RRH-UE mappings, which can be adequately designed to support applications (e.g., enhanced mobile broadband

(eMBB), ultra-reliable low-latency communications (URLLC) and massive machine type communications (mMTC)) that require reserved network services, while heading towards network greenness.

Cross-References

- [Network Slicing-Enabled Green C-RAN](#)
- [Energy Efficiency of Cellular Networks](#)
- [Network Slicing-Enabled Green C-RAN](#)

References

- Alimi IA, Teixeira AL, Monteiro PP (2018) Toward an efficient C-RAN optical fronthaul for the future networks: a tutorial on technologies, requirements, challenges, and solutions. *IEEE Commun Surv Tutor* 20(1):708–769
- Aqeeli E, Moubayed A, Shami A (2018) Power-aware optimized RRH to BBU allocation in C-RAN. *IEEE Trans Wireless Commun* 17(2):1311–1322
- Arnold O et al (2010) Power consumption modeling of different base station types in heterogeneous cellular networks. In: *Proceedings of the future network and mobile summit*
- Ashraf I, Boccardi F, Ho L (2011) SLEEP mode techniques for small cell deployments. *IEEE Commun Mag* 49(8):72–79
- Auer G et al (2011) How much energy is needed to run a wireless network? *IEEE Wireless Commun* 18(5):40–49
- Checko A et al (2015) Cloud RAN for mobile networks—a technology overview. *IEEE Commun Surv Tutor* 17(1):405–426
- Chiang Y, Liao W (2017) Multicell sleeping control and transmit power adaptation in green heterogeneous networks. In: *Proceedings of the IEEE GLOBECOM*
- Cisco (2019) Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. <https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-738429.pdf>
- Foukas X et al (2017) Network slicing in 5G: survey and challenges. *IEEE Commun Mag* 55(5):94–100
- Holtkamp H et al (2014) Minimizing base station power consumption. *IEEE J Selected Areas Commun* 32(2):297–306
- I C et al (2014) Toward green and soft: a 5G perspective. *IEEE Commun Mag* 52(2):66–73
- Rost P et al (2017) Network slicing to enable scalability and flexibility in 5G mobile networks. *IEEE Commun Mag* 55(5):72–79
- Saxena N, Roy A, Kim H (2016) Traffic-aware cloud RAN: a key for green 5G networks. *IEEE J Select Areas Commun* 34(4):1010–1021
- Wang K, Zhou W, Mao S (2017) On joint BBU/RRH resource allocation in heterogeneous cloud-RANs. *IEEE Internet Things J* 4(3):749–759
- Yao J, Ansari N (2018) QoS-aware joint BBU-RRH mapping and user association in cloud-RANs. *IEEE Trans Green Commun Netw* 2(4):881–889
- Zhang H et al (2017) Network slicing based 5G and future mobile networks: mobility, resource management, and challenges. *IEEE Commun Mag* 55(8):138–145
- Zhou X et al (2016) Network slicing as a service: enabling enterprises' own software-defined cellular networks. *IEEE Commun Mag* 54(7):146–153

Network Stack for Wireless Sensor Networks

- [Protocol Stack Architecture for Wireless Sensor Networks](#)

Network Transport Capacity

- [Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space](#)

Network Virtualization in the Data Center

- [Data Center Network Virtualization](#)

Networking of Marine Vessels

- [Maritime Internet of Vessels](#)

Network-Layer Mobility Management

Sandra Céspedes
Department of Electrical Engineering,
Universidad de Chile, Santiago, Chile

Synonyms

[IP mobility management](#)

Definitions

Network-layer mobility management refers to the ability of a mobile node to maintain the IP address while it moves. In traditional IP net-

works, IP addresses are employed at the same time as host identifiers and locators. The former means that transport protocols in a TCP/IP stack operate based on a combination of IP addresses and ports as the identities for a communication between two processes. The latter means that IP addresses are also used as an aggregatable system of addresses that help to route packets toward their destination. The impact of mobility across different IP domains is observed in Fig. 1, where a mobile node has established a communication with an application server. Once the mobile node changes its IP address from MN IP1 to MN IP2, packets sent by the application server with destination IP address MN IP1 get lost in the network (i.e., IP address unreachable). In addition, the translation in the DNS system gets outdated due to the change of IP address of the mobile node; hence, if a communication is to be established from an external server to the mobile node, it will fail until the entry in the DNS system is updated.

A network-layer mobility management protocol intends to keep the same IP address of the mobile node even when the mobile node changes the point of attachment when moving among different IP networks. Therefore, the change of IP network becomes transparent to active applications and transport protocols running on top of IP.

Historical Background

A mobility management mechanism is in charge to handle two aspects: the location of the mobile device and the maintenance of the connections while the mobile user is in motion (Akyildiz et al. 1999). In general, when mobility support is provided with a mobility management protocol, the network is able to locate their mobile users, so that new connections directed to those users can be successfully established. Once a connection has been established, it should remain active while the user is moving to a new point of attachment, and the transition to the new service area should be as transparent to the user as possible.

Mobility management could be performed at different layers on the stack of communication protocols. For the specific case of network-layer

mobility management, it is related to protocols in charge to handle the IP assignment in order to provide IP address continuity when the node changes the point of attachment to the network.

As illustrated in Fig. 2, according to the domain of action, mobility management protocols provide global mobility when IP address continuity is enabled although the movement of the mobile node causes a change to the global, end-to-end routing of packets, including a change to the global IP address (Kempf 2007). Localized mobility management protocols provide IP address continuity under a single administrative domain (and typically, a single access network).

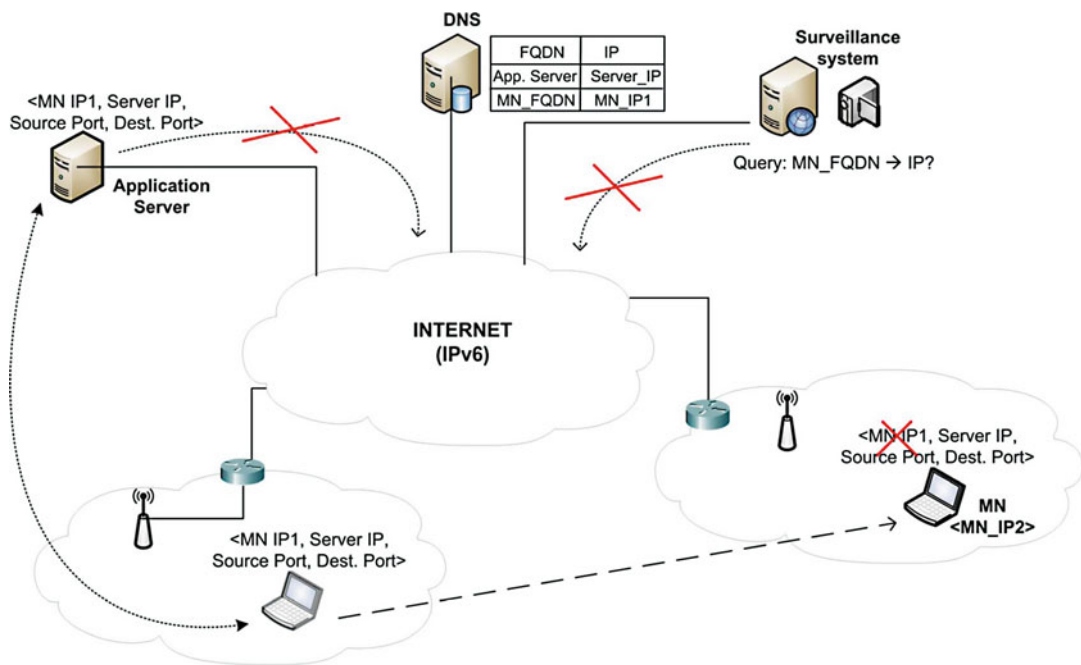
The first version of a network-layer mobility management protocol for IP networks was standardized at the IETF in the mid-1990s. The protocol, namely, IP Mobility Support – RFC 2002 (Perkins 2002) – was first standardized for IPv4 and later on for IPv6. Several extensions and variations in hierarchical networks, as well as updated versions, have been standardized ever since (cf. Key Proposals).

Key Proposals

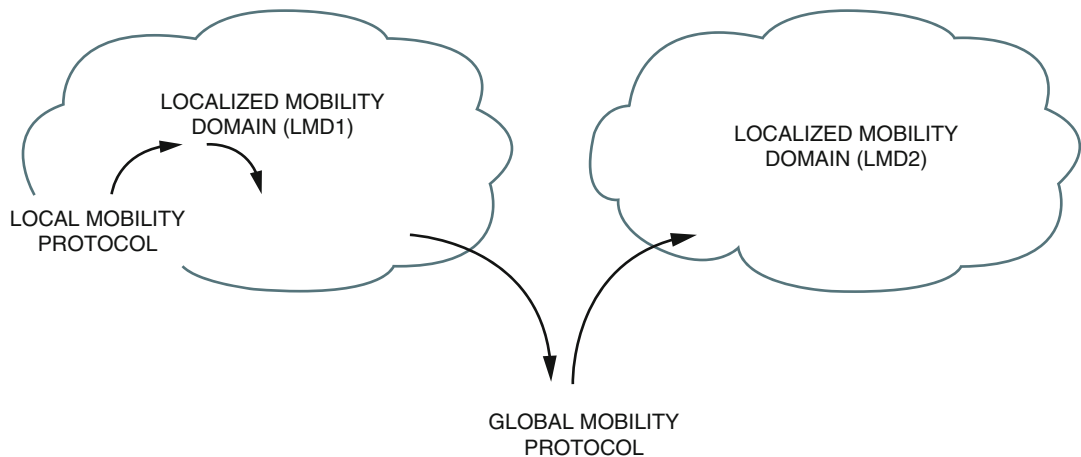
In the most traditional form of classification, network-layer mobility management protocols have been categorized as host-based or network-based. The division refers to the entity responsible for detecting and managing mobility to maintain the IP address assignment. In the case of host-based mobility, it is the mobile node's responsibility, whereas in the case of network-based mobility, the network is in charge of detecting the node's movement and exchange all the signalling necessary to provide IP address continuity.

Host-Based Mobility Protocols

Among the category of host-based mobility protocols are the protocols Mobile IP and the Network Mobility (NEMO) Basic Support. Mobility support in IPv6 (MIPv6) was initially defined by the IETF in 2004, with an updated version being released in 2011 (Perkins et al.



Network-Layer Mobility Management, Fig. 1 Illustration of the effects of mobility in IP-based communications



Network-Layer Mobility Management, Fig. 2 Global and localized mobility management

2011). In general, MIPv6 enables the mobile node to keep using its home address (i.e., the IP address assigned in its home link), even when the node moves to a visited network. When the mobile node moves to a visited network, it configures an IP address, known as the care-of-address, through conventional IPv6 mechanisms such as stateless or stateful auto-configuration. This care-of-address allows the node to establish

communications at the new location, but it also serves for establishing an association, or “binding,” between the mobile node’s home address and the care-of-address. The binding is performed when the mobile node is away from the home network. At that point, the mobile node registers the newly acquired care-of-address with a router, known as the home agent, in its home network. Therefore, if a correspondent

node sends packets to the mobile node, they are first routed to the mobile node's home network, where the home agent encapsulates each packet with a new IP header and redirects them to the visited network (i.e., the extra header indicates the mobile node's care-of-address as the new destination). Then, the mobile node decapsulates and processes the original IP packets.

An optimized version to avoid the pass through the home agent has been also defined for MIPv6 (Arkko et al. 2007). If the correspondent node supports mobility, then the mobile node can inform about the new location directly to the correspondent node. As a result, packets coming from the correspondent node are routed directly to the visited network, with no need to be sent first to the home agent. This enhancement also requires a security procedure, called Return Routability, which permits the mobile node to prove that the claimed home address and care-of-address are indeed assigned to it, preventing in this way man-in-the-middle attacks.

In the case of NEMO Basic Support (Devarapalli et al. 2005), it is an extension to Mobile IP for the support of mobile networks instead of single hosts. Nodes in the mobile network are served by a mobile router, and they configure IP addresses from a mobile network prefix advertised by the mobile router. When the mobile router connects to an access router in a visited network, it acquires a care-of-address and establishes a tunnel with the home agent, similar to MIPv6. In this way, such tunnel is used for communications of the mobile network nodes with any correspondent node. Although the standard does not include an optimized version of NEMO BS, extensive research has been done to improve the performance of the protocol in terms of delay and throughput (Petander et al. 2006).

Network-Based Mobility Protocols

Proxy Mobile IPv6 (Gundavelli et al. 2008) is a network-based mobility management approach in which the network, on behalf of the mobile node, performs all the signalling required to provide IP mobility. An entity named the mobile access gateway detects new connections and exchanges proxy binding updates and proxy

binding acknowledgements (PBU/PBA) with a centralized entity known as the local mobility anchor (LMA). The LMA is a manager for network prefixes assigned to nodes inside the administrative domain. When a handover occurs, the new mobile access gateway notifies the new connection to the LMA (i.e., it sends a PBU). Then, the LMA identifies the MN and assigns the same network prefix to it (i.e., it replies with a PBA). In this way, every time the mobile node roams inside the PMIP domain, the node does not detect any changes at the network layer, and the mobility is transparent to upper layers in the stack of protocols.

Future Directions

A common characteristic of traditional IP mobility management protocols is the use of a centrally deployed mobility anchor in charge of managing traffic from millions of mobile nodes, which poses several problems since a small number of mobility anchors end up managing both mobility and reachability for mobile nodes. Since 2012, a new initiative at the IETF, namely, the Distributed Mobility Management – DMM – working group, proposes a distributed mobility management architecture and explores deployment practices of existing mobility protocols to satisfy distributed mobility management requirements (Liu et al. 2015). The working group considers solutions based on physical or virtualized networking functions and the possibility for a mobile node to select between the IP address in the visited network (i.e., care-of address) and the “permanent” IP address (i.e., home address) depending on the specific requirements of the application.

Cross-References

- [Distributed IP Mobility Management](#)
- [Mobile IP](#)
- [Proxy Mobile IPv6](#)

References

- Akyildiz IF, McNair J, Ho J, Uzunalioglu H, Wang W (1999) Mobility management in next-generation wireless systems. *Proc IEEE* 87(8):1347–1384
- Arkko J, Vogt C, Haddad W (2007) Enhanced route optimization for mobile IPv6. IETF RFC 4866
- Devarapalli V, Wakikawa R, Petrescu A, Thubert P (2005) Network mobility (nemo) basic support protocol. IETF Secretariat, RFC 3963
- Gundavelli A, Leung K, Devarapalli V, Chowdhury K, Patil B (2008) Proxy mobile IPv6. IETF Secretariat, RFC 5213
- Kempf J (2007) Problem statement for network-based localized mobility management (NETLMM). IETF RFC 4830
- Liu D, Zuniga J, Seite P, Chan H, Bernardos C (2015) Distributed mobility management: current practices and gap analysis. IETF Secretariat, RFC 7429
- Perkins C (2002) IP mobility support. IETF Secretariat, RFC 2002
- Perkins C, Johnson D, Arkko J (2011) Mobility support in IPv6. IETF Secretariat, RFC 6275
- Petander H, Perera E, Lan KC, Seneviratne A (2006) Measuring and improving the performance of network mobility management in IPv6 networks. *IEEE J Sel Areas Commun* 24(9):1671–1681. <https://doi.org/10.1109/JSAC.2006.875116>

Neural Communication

► Neuronal Communication Channels

Neuronal Communication Channels

Hamideh Ramezani¹, Tooba Khan², Ergin Dinc¹, and Özgür Barış Akan¹

¹University of Cambridge, Cambridge, UK

²Koç University, Istanbul, Turkey

Synonyms

Neural communication; Neuronal signaling; Neuro-spike communication

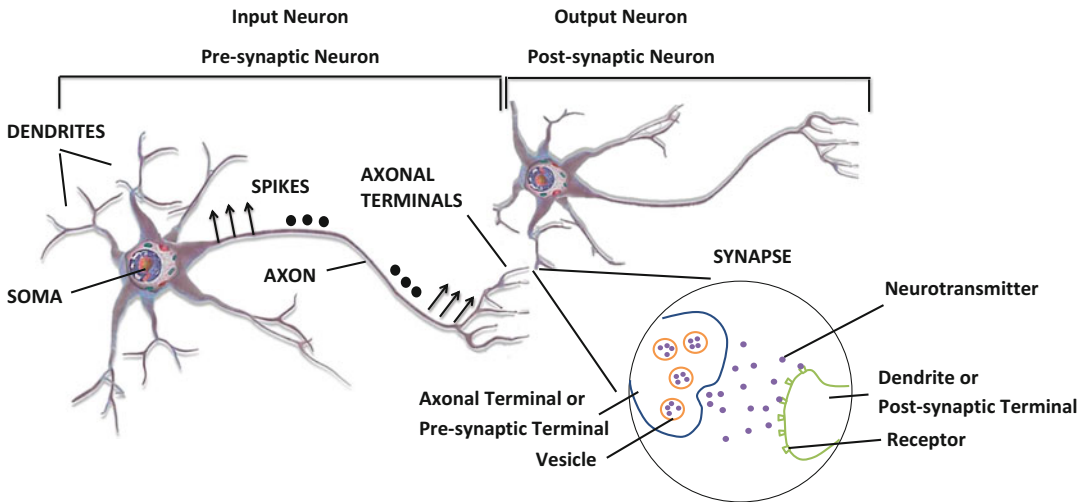
Definitions

Neuronal communication or *neuro-spike communication* (NSC) is the method neurons use to transmit information, which is encoded into *spike* trains. This communication starts with propagation of spikes through *axon* of the input neuron until reaching the *axonal terminals* located at the end of the axon. This initiates the fusion of *vesicles*, i.e., packets of *neurotransmitters*, with cell membrane resulting in the release of neurotransmitters to the gap between input and output neurons, called *synapse*. After diffusing through the synapse, neurotransmitters bind to the respective receptors located on the surface of the postsynaptic neuron, which leads to opening of ionic channels and changing the postsynaptic membrane potential. If these changes are strong enough, the output neuron detects spike arrival, i.e., generates a spike.

Historical Background

Neuronal signaling, a natural phenomenon enabling communication at nanoscale, advocates the idea of nanoscale communication networks. Thus, it is a highly investigated intrabody nanonetwork to understand the communication paradigm underlying information transfer between nerve cells. In NSC, sensory or first-order neurons encode information from outside world into spike trains. This information is proposed to be either encoded in exact timing of the spike or in the spiking rate over a certain period of time, known as temporal and rate coding, respectively (Bear et al. 2007). These spikes are transmitted through neuronal network to reach higher brain regions for processing or memory storage. In the following, the physiology of NSC is briefly explained. Single-input single-output (SISO) NSC is comprised of three parts as shown in Fig. 1.

Axonal propagation Axon is the nerve fiber that conducts electrical impulses, known as action potentials or spikes, away from the cell



Neuronal Communication Channels, Fig. 1 Neural anatomy of communication among a pair of neurons (Ramezani and Akan 2018a)

body. Some axons are covered with fat layer called myelin, which works as an electrical insulation layer and results in faster conduction of spikes (Bear et al. 2007). Spikes were considered as all-or-none phenomenon for a long time; however, depending on the previously propagated spikes through the axon, spike shape can change during axonal transmission (Geiger and Jonas 2000; Feng et al. 2014; Ramezani and Akan 2017).

Synaptic transmission Information from one neuron to another neuron is transferred over a synapse. The synapses are of two types: (i) electrical synapses, called gap junctions, which are electrical links formed between two neurons, and (ii) chemical synapses, where information is transmitted through neurotransmitters, i.e., a type of chemical messengers, across synaptic cleft. However, the chemical synapses are the most common ones (Bear et al. 2007). Thus, most of the NSC studies are focused on chemical synapses, and the word synapse is usually used to refer to a chemical synapse in these studies. In presynaptic terminal, the neurotransmitters are stored in vesicles that are divided into three pools, i.e., reserve pool, recycling pool, and readily releasable pool (RRP), depending on their availability to be released in response to a

spike (Rizzoli and Betz 2005). It takes the longest to mobilize the vesicles in the reserve pool, while in RRP, vesicles are ready to be mobilized upon arrival of a spike.

Spikes are propagated through the axon until reaching the presynaptic terminal, where they lead to opening of voltage-gated calcium channels (VGCCs) allowing the influx of calcium ions. Increase in calcium concentration inside the cell initiates the fusion of vesicles from RRP with the cell membrane releasing neurotransmitters into the synapse. Different types of neurotransmitters are used for different functionalities of the brain such as glutamate, dopamine, and serotonin. These neurotransmitters diffuse through the synaptic cleft to reach the terminal of the receiving neuron, known as postsynaptic terminal (Bear et al. 2007).

Spike generation Released neurotransmitters bind to their respective receptors present on the postsynaptic terminal. This binding opens ionic channels that allow influx or efflux of ions resulting in the change of the postsynaptic membrane potential. This membrane potential change can either be excitatory, i.e., facilitating the output neuron to fire a spike, or inhibitory, i.e., inhibiting the neuron from firing a spike. Thus, these are termed as excitatory postsynaptic

potential (EPSP) and inhibitory postsynaptic potential (IPSP), respectively (Bear et al. 2007).

In real scenario, a postsynaptic neuron has multiple synapses, i.e., it receives input from multiple presynaptic neurons, where each input can produce EPSP or IPSP at their respective synapses. However, for spike generation, the membrane potentials are integrated at soma. If this accumulated change is larger than spike generation threshold, which depends on the previous stimulations the neuron has received (Kobayashi et al. 2009), the postsynaptic neuron generates a spike.

ICT Methods for Neuro-Spike Communication (NSC)

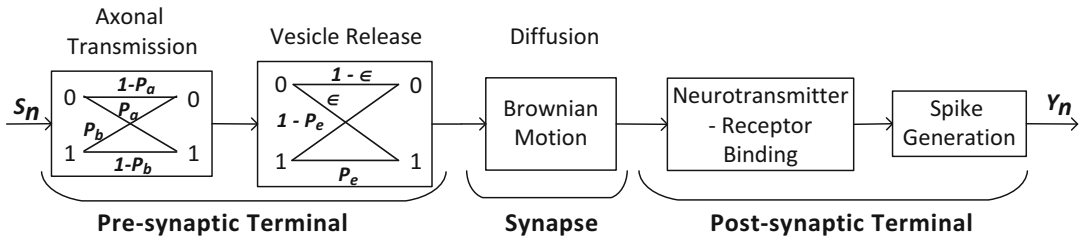
Engineering-based studies of neuronal signaling, which are from ICT point of view, aim to characterize different biological events involved in this communication to evaluate the communication theoretical performance of the brain and provide nanoscale communication paradigms (Balevi and Akan 2013; Malak and Akan 2013; Veletić et al. 2016a). In Manwani and Koch (2001), detection and estimation paradigms are used to calculate information transfer rate of a SISO NSC. In Balevi and Akan (2013), the performance of NSC is evaluated by deriving the probability of error detection in the output neuron. The rate region of synaptic communication is derived in Malak and Akan (2013) by investigating the multiple access to the output neuron. The same approach is used in Ramezani et al. (2017) to find the rate region of SISO NSC with multiple synaptic terminals between input and output neurons. An upper bound is derived for the capacity of SISO synaptic communication in Veletić et al. (2016b). The impact of spike generation on the performance of multiple-input single-output (MISO) NSC is investigated in Ramezani et al. (2018b). Recent studies have focused on the randomness in different processes involved in the NSC. The SISO NSC channel is studied in Maham (2015) under axonal noise. Moreover, the effects of variations in spike shape on the efficiency of synaptic communication are studied in Ramezani and

Akan (2018a) by modeling the gating of VGCCs and its impact on the release of neurotransmitters. Furthermore, the changes in the vesicle release probability with time due to unavailability of vesicles for release are considered in Ramezani and Akan (2018b) and Ramezani et al. (2018a) to find the capacity of vesicle release and synaptic communication, respectively. Results depicted that ignoring the variation in vesicle release probability leads to overestimation of the capacity of the NSC. Despite the recent efforts toward having a more accurate NSC model, the theoretical models still fail to comprehensively capture aspects of the neuron dynamics and their impact on the performance of this communication channel. Some of these aspects are (i) possible alterations in spike shape in healthy and disease-affected neurons, (ii) changes in synaptic strengths due to both short- and long-term plasticities, and (iii) variations in the spike generation threshold. In addition, experimental studies are needed to validate the accuracy of these theoretical models. In the remainder of this section, a modeling approach is explained for the NSC in which different sources of randomness are taken into account. In addition, the open challenges of this modeling approach are highlighted.

By discretizing time into windows of equal duration, i.e., Δt_s , where Δt_s is the duration of a spike (Ramezani and Akan 2018b), the block diagram of the SISO NSC is derived as shown in Fig. 2.

Input signal The input signal, i.e., S_n , indicated spike generation by input neuron in n th time slot, which is governed by Poisson process (Manwani and Koch 2001). Thus, the probability of having one spike at input is derived as $P\{S_n = 1\} = 1 - \exp(-\lambda \Delta t_s)$, where λ is the rate of the Poisson process.

Axonal propagation As a result of random opening of ionic channels, erroneous spikes might be added to the input spike train during axonal transmission (Maham 2015). Moreover, failures can happen during axonal propagation due to different diseases. Thus, in Fig. 2, P_a and P_b are used for the probability of erroneous



Neuronal Communication Channels, Fig. 2 Block diagram of NSC channel

spike generation and failure in spike propagation, respectively. Note that in this model, the spike is considered to be all or none, i.e., its shape does not change. However, variations in the spike shape during axonal transmission are recorded with the use of direct recording from axons (Jensen and Durand 2009; Sasaki 2013; Feng et al. 2014), which consequently affect the performance of NSC (Ramezani and Akan 2018a). Hence, further experimental studies are required to comprehensively distinguish source of the spike shape changes and their impact on the vesicle release process.

Vesicle release process Different models exist in the literature for vesicle release process including deterministic (Veletić et al. 2016a) and stochastic models, where the vesicle release process is modeled by a binary Z channel (Manwani and Koch 2001; Matveev and Wang 2000). Deterministic models fail to capture the trial-to-trial variability in the vesicle release process. In the stochastic models, the release probability is either considered to be fixed (Manwani and Koch 2001) or changing with time (Matveev and Wang 2000). The latter one is more accurate as vesicle release probability depends on the availability of vesicles for release, which changes with time. Utilizing the release model given in Matveev and Wang (2000), it is shown in Ramezani and Akan (2018b) that the probability of having N vesicles for release at the beginning of a time step can be derived based on a finite-state Markov chain by considering a pool-based vesicle release model. Then, the vesicle release probabilities P_e and ϵ can be derived according to the pool-based release model and the derived Markov chain.

Diffusion process Diffusion of neurotransmitters through the synapse follows Brownian motion. This is used in Khan et al. (2017) to derive the expected value of neurotransmitters concentration in the cleft t seconds after release by considering the synaptic geometries and neurotransmitter reuptake. Further studies are required to evaluate whether the randomness in diffusion process must be included in NSC model.

Neurotransmitter receptor binding If a neurotransmitter reaches to the vicinity of a post-synaptic receptor, it can bind to it with some probability, which depends on type of the receptor. As shown in Ramezani et al. (2018a), the probability of binding can be derived by utilizing the kinetic model of neurotransmitter receptor binding and the concentration of neurotransmitters near receptors.

Spike generation A simple spike generation model is derived by comparing the postsynaptic membrane potential with a constant threshold (Ramezani et al. 2018b). Accurate models must encompass the variations in the spiking threshold by time, for instance, by utilizing a multi-timescale adaptive threshold (Kobayashi et al. 2009).

Key Applications

The study of neural signaling from communication theoretical perspective will pave the way for development of bio-inspired nanonetworks for various applications requiring biocompatible net-

work paradigm as well as developing ICT-based treatment and diagnosis techniques for neurodegenerative diseases.

Nanoscale Transceivers and Artificial Synapses

ICT analysis of NSC can be exploited to implement artificial devices providing two-way communication in nanoscale. This enables the development of bio-inspired nanonetworks for various applications from biomedical to industrial and environmental fields (Akyildiz et al. 2008). Moreover, it can enable the technologies required for the realization of Internet of Bio-Nano Things (IoBNT) (Akyildiz et al. 2015). This gives rise to vast applications in healthcare including remote monitoring of patient's vital signs through the Internet, administering periodic dosage of drugs in patients, and repairing communication failure between vital organs and cells within the body.

Another application of these devices is communicating with neurons as artificial synapses, i.e., stimulating a neuron and detecting neural signals. For this purpose, there are three main methods to interact with neurons: electrical, chemical, and optical methods. First, the patch clamp method has been introduced by Sakmann and Neher (1984) to study electrophysiological activities in electrically excitable cells such as neurons and cardiomyocytes. Since then, the patch clamp method is the gold standard in electrophysiology. Inspired by this technique, high-density microelectrode arrays have been introduced to stimulate/detect neuro-spikes from lower number of neurons, e.g., biocompatible and transparent graphene electrodes are utilized for neural signaling in Kuzum et al. (2014). Electrical signaling can achieve high resolution with high number of electrodes; however, continuous electrical stimulation may cause stress in the target cell.

The second method is chemical stimulation in the form of neurotransmitters, which mimics the natural way of communication in chemical synapses. Thus, chemical stimulation does not create cell stress as in the electrical stimulation, but the interference is a significant chal-

lenge due to the diffusion of neurotransmitters to neurons that are not targeted. A hybrid device based on PEDOT:PSS electrodes, which is capable of chemical stimulation and electrical detection, is proposed for lowering the neural activity to control epileptiform activity by Jonsson et al. (2016). Chemical stimulation can be also exploited for neural stimulation by using excitatory neurotransmitters. Furthermore, optogenetics is a popular method for stimulating neurons. Engineered photoactive neurons are stimulated via specific wavelengths to create excitatory or inhibitory effects (Wirdatmadja et al. 2017). Optical stimulation provides accurate stimulation of target neuron with less noise, but requires biological modification of the neuron. At the end, these techniques have some advantages/disadvantages, and the current trend is to combine different techniques to exploit the advantages, e.g., a device having six electrodes, one optical waveguide, and two microfluidic channels was proposed by Park et al. (2017).

In addition to the interfacing with neurons, synaptic plasticity, i.e., the change in synaptic communication over time, can be artificially implemented by novel memory devices, i.e., memristive devices whose resistance depends on the previous currents (Alibart et al. 2013). Thus, memristive devices can mimic a synapse to realize long-term potentiation and depression of synapses. Memristive devices together with artificial synapses also pave the way for bio-inspired neuromorphic computation devices.

Diagnosis and Treatment of Neuronal Dysfunction

Most of the neurodegenerative diseases are the result of either abnormal interaction or loss of communication between neurons. For instance, in Alzheimer's disease, the capacity of brain to store and retrieve memories is degraded due to the loss of communication between neurons, i.e., due to loss of synapses (Sheng et al. 2012). Destruction of myelin sheath over axon results in multiple sclerosis, which reduces the information transfer rate through the affected network (Steinman 1996). Parkinson's disease is caused by falling dopamine levels in the brain, which

controls the vesicle release in neurons, which are responsible for information transfer between neurons (Harnois and Di Paolo 1990). Amyotrophic lateral sclerosis (ALS) causes death of motor neurons disconnecting the brain from the peripheral nervous nanonetwork. This is analogous to communication link failure in a network affecting the control of a brain on essential muscle activities (Leigh and Ray-Chaudhuri 1994). Another type of communication failure may occur due to spinal cord injury. Thus, unveiling the ICT-based characteristics of the disease-affected and healthy neuronal networks may enable the neurologists to diagnose diseases by detecting fundamental differences between the two networks.

Traditional treatments for neurodegenerative diseases and spinal cord injuries include medicine and rehabilitation techniques. The ICT-based understanding of nervous nanonetworks may also enable the development of new treatment techniques such as building artificial neural implants to cure communication failure in the nervous system. Advancements in this direction have been published in Capogrosso et al. (2016), where researchers have achieved leg movement in a paralyzed nonhuman primate by using implants in the motor cortex and spinal cord that communicate through wireless channel, bypassing the lesion in spinal cord. Neural interfacing also enables artificial retinas to restore loss of vision, artificial cochlea to restore loss of hearing, and artificial nose to restore loss of smell by converting sensory data to electrical signal that can be received by target neurons (Dobelle 2000; Eddington et al. 1978).

Conclusions and Future Research Challenges

In this article, an overview of neural signaling is provided from an ICT perspective including its various applications such as artificial synapses to interact with the nervous system in order to develop a bio-inspired nanoscale communication paradigm and ICT-inspired diagnosis and treatment methods for neuronal dysfunctions.

There are several open research challenges to be tackled for the ICT-based analysis of neural signaling and possible applications of these techniques. As molecular communication and nanonetworks being an emerging research field, there exist numerous theoretical studies on ICT-based investigation of neural signaling based on physiological data. Theoretical models often fail to capture the action potential shape variation, spike-timing-dependent plasticity, and synaptic long-/short-term plasticity. For this reason, the development of realistic neural signaling model supported by experiments stands as an important open research problem.

Artificial synapses also include many open research issues and implementation problems. The first issue is reaching the ultimate neural interfacing resolution, i.e., one-to-one pairing between electrodes and neurons. There exist some methods to interface with single neuron, but these methods have scalability problems, and many applications require interfacing with 100s or 1000s of neurons with the same device with low interference and noise. As these artificial devices need to operate inside living entities, i.e., body and brain of human/animal, biocompatibility of the devices is another significant problem. To this end, modern neural interfaces utilize novel biocompatible materials, e.g., PEDOT:PSS (Jonsson et al. 2016) and graphene (Kuzum et al. 2014). The lifetime of the devices is critical as well. There are two important factors that limit the lifetime of such devices. First, the device may lose its sensing/stimulating properties due to slowly dissolving properties of materials that are used. In addition, battery-powered devices offer limited lifetime that is prone to random depletions, and changing the batteries of invasive artificial synapses is often impossible. Thus, artificial synapses are required to have energy-harvesting techniques to empower its operations. For this purpose, wireless powered optogenetic devices with RF energy-harvesting capabilities are proposed by Wirdatmadja et al. (2017). Novel energy-harvesting techniques, e.g., energy harvesting from molecules, are required for continuous operation of artificial in-body devices.

Cross-References

- [Applications of Molecular Communication Systems](#)
- [Brain-Machine Interfaces](#)
- [Brownian Motion](#)
- [Connectivity via Molecular Signaling](#)
- [Nanonetworks](#)

References

- Akyildiz IF, Brunetti F, Blázquez C (2008) Nanonetworks: a new communication paradigm. *Comput Netw* 52(12):2260–2279
- Akyildiz I, Pierobon M, Balasubramaniam S, Koucheryavy Y (2015) The internet of bio-nano things. *IEEE Commun Mag* 53(3):32–40
- Alibart F, Zamanidoost E, Strukov DB (2013) Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nat Commun* 4:2072
- Balevi E, Akan OB (2013) A physical channel model for nanoscale neuro-spike communications. *IEEE Trans Commun* 61(3):1178–1187
- Bear MF, Connors BW, Paradiso MA (2007) *Neuroscience*, vol 2. Lippincott Williams & Wilkins, Boston
- Capogrosso M, Milekovic T, Borton D, Wagner F, Moraud EM, Mignardot JB, Buse N, Gandar J, Barraud Q, Xing D, Rey E, Duis S, Jianzhong Y, Ko WKD, Li Q, Detemple P, Denison T, Micera S, Bezard E, Bloch J, Courtine G (2016) A brain-spine interface alleviating gait deficits after spinal cord injury in primates. *Nature* 539(7628):284–288
- Dobelle WH (2000) Artificial vision for the blind by connecting a television camera to the visual cortex. *ASAIO J* 46(1):3–9
- Eddington DK, Dobelle WH, Brackmann DE (1978) Auditory prostheses research with multiple channel intracochlear stimulation in man. *Ann Otol Rhinol Laryngol* 87(6 II SUPP. 53):1–39
- Feng Z, Yu Y, Guo Z, Cao J, Durand DM (2014) High frequency stimulation extends the refractory period and generates axonal block in the rat hippocampus. *Brain Stimul* 7(5):680–689
- Geiger JR, Jonas P (2000) Dynamic control of presynaptic Ca^{2+} inflow by fast-inactivating K^{+} channels in hippocampal mossy fiber boutons. *Neuron* 28(3):927–939
- Harnois C, Di Paolo T (1990) Decreased dopamine in the retinas of patients with Parkinson's disease. *Invest Ophthalmol Vis Sci* 31(11):2473–2475
- Jensen AL, Durand DM (2009) High frequency stimulation can block axonal conduction. *Exp Neurol* 220(1):57–70
- Jonsson A, Inal S, Ilke Uguz, Williamson AJ, Kergoat L, Rivnay J, Khodagholy D, Berggren M, Bernard C, Malliaras GG, Simon DT (2016) Bioelectronic neural pixel: chemical stimulation and electrical sensing at the same site. *Proc Natl Acad Sci* 113(34):9440–9445
- Khan T, Bilgin BA, Akan OB (2017) Diffusion-based model for synaptic molecular communication channel. *IEEE Trans Nanobiosci* 16(4):299–308
- Kobayashi R, Tsubo Y, Shinomoto S (2009) Made-to-order spiking neuron model equipped with a multi-timescale adaptive threshold. *Front Comput Neurosci* 3:9
- Kuzum D, Takano H, Shim E, Reed JC, Juul H, Richardson AG, De Vries J, Bink H, Dichter MA, Lucas TH, Coulter DA, Cubukcu E, Litt B (2014) Transparent and flexible low noise graphene electrodes for simultaneous electrophysiology and neuroimaging. *Nat Commun* 5:5259
- Leigh PN, Ray-Chaudhuri K (1994) Motor neuron disease. *J Neurol Neurosurg Psychiatry* 57(8):886
- Maham B (2015) A communication theoretic analysis of synaptic channels under axonal noise. *IEEE Commun Lett* 19(11):1901–1904
- Malak D, Akan OB (2013) A communication theoretic analysis of synaptic multiple-access channel in hippocampal-cortical neurons. *IEEE Trans Commun* 61(6):2457–2467
- Manwani A, Koch C (2001) Detecting and estimating signals over noisy and unreliable synapses: information-theoretic analysis. *Neural Comput* 13(1):1–33
- Matveev V, Wang XJ (2000) Implications of all-or-none synaptic transmission and short-term depression beyond vesicle depletion: a computational study. *J Neurosci* 20(4):1575–1588
- Park S, Guo Y, Jia X, Choe HK, Grena B, Kang J, Park J, Lu C, Canales A, Chen R, Yim YS, Choi GB, Fink Y, Anikeeva P (2017) One-step optogenetics with multifunctional flexible polymer fibers. *Nat Neurosci* 20(4):612–619
- Ramezani H, Akan OB (2017) A communication theoretic modeling of axonal propagation in hippocampal pyramidal neurons. *IEEE Trans Nanobiosci* 16(4):248–256
- Ramezani H, Akan OB (2018a) Impacts of spike shape variations on synaptic communication. *IEEE Trans Nanobiosci* 17(3):260–271
- Ramezani H, Akan OB (2018b) Information capacity of vesicle release in neuro-spike communication. *IEEE Commun Lett* 22(1):41–44
- Ramezani H, Koca C, Akan OB (2017) Rate region analysis of multi-terminal neuronal nanoscale molecular communication channel. In: *Proceedings of IEEE Nano'17, Pittsburgh*
- Ramezani H, Khan T, Akan OB (2018a) Information theoretic analysis of synaptic communication for nanonetworks. In: *Proceedings of 37th IEEE INFOCOM, Honolulu. IEEE*

- Ramezani H, Khan T, Akan OB (2018b) Sum rate of miso neuro-spike communication channel with constant spiking threshold. *IEEE Trans Nanobiosci* 17(3): 342–351
- Rizzoli SO, Betz WJ (2005) Synaptic vesicle pools. *Nat Rev Neurosci* 6(1):57
- Sakmann B, Neher E (1984) Patch clamp techniques for studying ionic channels in excitable membranes. *Ann Rev Physiol* 46(1):455–472
- Sasaki T (2013) The axon as a unique computational unit in neurons. *Neurosci Res* 75(2):83–88
- Sheng M, Sabatini BL, Südhof TC (2012) Synapses and Alzheimers disease. *Cold Spring Harb Perspect Biol* 4(5):a005777
- Steinman L (1996) Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system. *Cell* 85(3):299–302
- Veletić M, Floor PA, Babić Z, Balasingham I (2016a) Peer-to-peer communication in neuronal nanonetwork. *IEEE Trans Commun* 64(3):1153–1166
- Veletić M, Floor PA, Chahibi Y, Balasingham I (2016b) On the upper bound of the information capacity in neuronal synapses. *IEEE Trans Commun* 64(12):5025–5036
- Wirdatmadja SA, Barros MT, Koucheryavy Y, Jornet JM, Balasubramaniam S (2017) Wireless optogenetic nanonetworks for brain stimulation: device model and charging protocols. *IEEE Trans Nanobiosci* 16(8): 859–872

Neuronal Signaling

- [Neuronal Communication Channels](#)

Neuro-spike Communication

- [Neuronal Communication Channels](#)

New Radio (NR)

- [5G Wireless](#)

New Variants of Mirai and Analysis

Zhen Ling¹, Yiling Xu¹, Yier Jin², Cliff Zou³, and Xinwen Fu^{4,5}

¹School of Computer Science and Engineering, Southeast University, Nanjing, China

²University of Florida, Gainesville, FL, USA

³University of Central Florida, Orlando, FL, USA

⁴University of Massachusetts Lowell, Lowell, MA, USA

⁵Department of EECS, University of Central Florida, Orlando, FL, USA

Synonyms

[Botnet](#); [DDoS](#); [IoT](#); [Malware](#)

Definition

Mirai, which means “the future” in Japanese, is a self-propagating malware that infects vulnerable networked IoT devices so as to turn them into part of a botnet that can be leveraged to perform large-scale distributed denial-of-service (DDoS) attacks.

Historical Background

The Mirai botnet was first discovered in August 2016 (Mal [2016](#)). It has been used in massive DDoS attacks, including an attack on KrebsOnSecurity in September 2016 which exceeded 600 Gbps (Krebs [2016](#)), an attack on OVH in September 2016 which exceeded 1 Tbps (Klaba [2016](#)), and an attack on Dyn in October 2016 (Hilton [2016](#)) resulting in the cripple of some well-known websites such as GitHub, Twitter, Reddit, Netflix, Airbnb, and many others (Williams [2016](#)). At its peak, Mirai infected 600k vulnerable IoT devices (Antonakakis et al. [2017](#)), whose IP addresses were spotted in 164 countries (Herzberg et al. [2016](#)). Since the source code for

Mirai was published on hackforums.net (Annasenpai 2016), the techniques can be digged deeply and may be adapted in other malware projects. The original creators of Mirai (i.e., Paras Jha, Josiah White, and Dalton Norman) has pleaded guilty in December 2017 (Department of Justice 2017) and sentenced to probation in September 2018 (Department of Justice 2018).

Foundations

Overview

In this chapter, we first present our analysis of the released source code of the Mirai malware for its architecture, scanning, and prorogation strategy (Antonakakis et al. 2017; Kambourakis et al. 2017; Ling et al. 2018). We then discuss why Mirai did not get attention in its prorogation phase until it deployed the DDoS attack. Finally, we present a new variant of Mirai that can improve the propagation speed. The goal of the study is to understand the impact of Mirai as both a worm and botnet.

Mirai Botnet

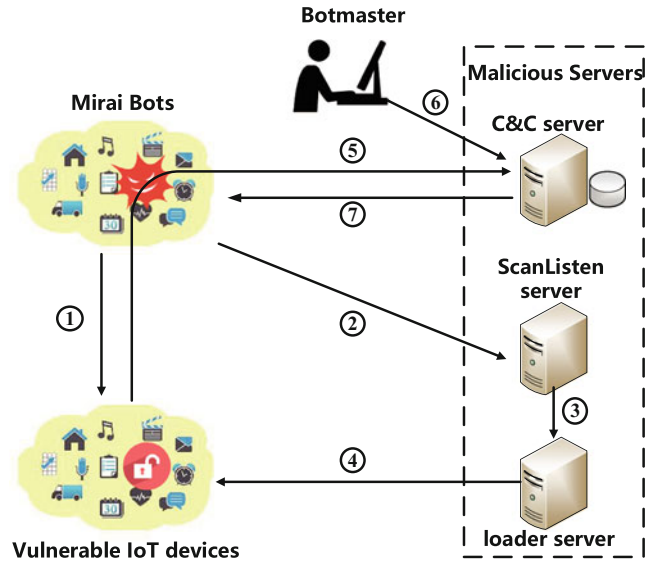
The Mirai botnet comprises four components as shown in Fig. 1: bots, a C&C (command and control) server, a scanListen server, and loader servers. The bots are a group of hijacked IoT devices via the Mirai malware. Starting with a scanning procedure on the port of the telnet service, the bots continue to perform a password dictionary attack using the username/password list already hardcoded in the Mirai malware. In this way, some vulnerable IoT devices are infected. The C&C server monitors the status of the botnet and controls the bots by obeying the commands from a botmaster, e.g., sending attack commands. On receiving the attack commands, the bots deploy various attacks such as distributed deny-of-service (DDoS) attacks. The scanListen server receives information of those newly discovered vulnerable IoT devices and relays it to the loader server. The loader server logs into a vulnerable device and infects it by executing the Mirai malware carried by the loader server. The IoT devices can be classified into three types:

(1) IoT devices that close the telnet service port, i.e., 23/2323 port. They are not accessible to the bots since the TCP connections are cut off permanently; (2) IoT devices that open the telnet service port but their username/password is absent in the username/password list of Mirai. The bots can perform the brute force attack against them but can never successfully compromise them; (3) IoT devices that open the telnet service port and their username/password is available in the list of Mirai. These devices can be compromised. Figure 1 illustrates the propagation procedure of Mirai. The step-by-step process is explained as follows.

- **Step 1 – Bots scanning IoT devices:** The propagation starts with a single IP, which is treated as a bot to ensure its consistency. The bot scans and detects vulnerable IoT devices over the Internet. The bot selectively probes the 23/2323 port of public IP addresses through TCP SYN scans. Note that some IP addresses are whitelisted by Mirai, including those of the US Postal Service, the Department of Defense, the Internet Assigned Numbers Authority (IANA), and IP ranges belonging to Hewlett-Packard and General Electric. Once a TCP SYN/ACK packet is acquired from an IoT device, a TCP connection is to be established between the bot and the IP address of the device. In order to log into the telnet server, the bot randomly selects a pair of usernames/password from its dictionary and tries it out to check whether the login is successful. By default, the dictionary has altogether 62 pairs of usernames/password. However, the password cracking trials are no more than ten times for a target.
- **Step 2 – Bots reporting information to scanListen server:** After a bot compromises a vulnerable IoT device, it reports the information of the device to the scanListen server, including the IP address, the port, and the username/password of the target. The domain name and port of scanListen server are also hardcoded into the malware.

New Variants of Mirai and Analysis, Fig. 1

Propagation process of Mirai



- **Step 3 – scanListen server relaying information to loader server:** The scanListen server receives the information of a compromised IoT device from the bot and then forwards it to the loader server.
- **Step 4 – Loader server installing Mirai:** The loader server first builds a TCP connection with the compromised device and then logs onto the device with the cracked username/password. After its login, the loader server starts detecting file transmission tools like *wget* or *ftp* on the target system. If a transmission tool is available, the Mirai malware is directly downloaded from the loader server and executed. Otherwise, the loader server connects to the telnet port of the target, reads a tiny binary file whose function is similar to that of *wget*, and writes *echo* into the target device. The hex characters from the binary file constitute the string for *echo*. Then the output stream of *echo* is redirected into a file and saved at the target device. Therefore, the binary file is transmitted from the loader server to the target. By executing the file, the Mirai malware can be downloaded from the loader server. Once the Mirai is executed, it shuts down Telnet, SSH, and HTTP server on the compromised device to avoid the login of other malwares or administrators. Mirai also

binds to port 48101 for the assurance that only one instance of the malwares is running on this device.

- **Step 5 – Bot registering with C&C server:** Once the Mirai malware is installed and executed on a vulnerable device, the device builds a TCP connection, registers with the C&C server, and gets involved in the Mirai botnet. The domain name and port of the C&C server are hardcoded into the malware. Then the bot performs Step 1.
- **Step 6 – Botmaster commanding bots to conduct attacks:** A botmaster can command the bots to perform attacks by sending instructions via the C&C server. The Mirai malware can be activated to launch various attacks like UDP DoS attack, TCP DoS attack, HTTP attack, and GREP attack.
- **Step 7 – Botnet attacking a target:** Once the C&C server receives the attacking instructions from the botmaster, it conveys the commands to the bots. Finally, the bots perform attacks against a target.

Scanning Strategy

The scanning algorithm of the released source code of the Mirai is worth studying since it is the key to its propagation speed. Therefore, we thoroughly analyze the Mirai scanning

Algorithm 1: Mirai Scanning Algorithm

Require: (a) μ - the number of SYN probes sent by Mirai each time, (b) m - the size of a table, (c) <i>table</i> - TCP connection table, (d) <i>table[i]</i> - store the <i>i</i>th TCP socket in <i>table</i>, (e) <i>table[i].state</i> - TCP connection state of the <i>i</i>th entry of <i>table</i>, e.g., connecting or connected. Ensure: Discover the vulnerable devices <pre> 1: while TRUE do 2: if time interval since last SYN probe $\geq 1s$ then 3: Send μ probes using a raw socket 4: end if 5: while TRUE do 6: Read SYN+ACKs packets from the raw socket 7: if error no data for reading no entries in table then 8: break 9: else if the <i>i</i>th entry of <i>table</i> is empty then 10: Build a connection, table[i].state = connecting 11: end if 12: end while 13: for $i = 1 : m$ do 14: if the connection in table[i] times out then 15: if table[i].state == connected then 16: re-connect(try times < 10), otherwise free table[i] 17: else if table[i].state == connecting then 18: free table[i] 19: end if 20: else if table[i].state == connecting then 21: Set the table[i] in write file descriptor set (wr fdset) 22: else if table[i].state == connected then </pre>	<pre> 23: Set the table[i] in read file descriptor set (rd fdset) 24: end if 25: end for 26: Select the write and read file descriptors 27: for $i = 1 : m$ do 28: if table[i] in wr fdset then 29: if connection error for table[i] then 30: free table[i] 31: else if connection success for table[i] then 32: table[i].state = connected 33: end if 34: end if 35: if table[i] in rd fdset then 36: while TRUE do 37: if connection error for table[i] then 38: re-connect(try times < 10), otherwise free table[i] 39: break 40: else if no data for reading then 41: break 42: else if data for the last command then 43: if login success then 44: report to scanListen server 45: else if login fail then 46: re-connect(try times < 10), otherwise free table[i] 47: end if 48: else 49: send one command 50: end if 51: end while 52: end if 53: end for 54: end while </pre>
---	---

algorithm in Algorithm 1. It is discovered that Mirai malware uses a uniform scanning strategy. Specifically, Mirai randomly scans public IP addresses and then randomly selects a pair of usernames/password from a hardcoded list for the dictionary attack. The TCP SYN scanning technique is employed to probe the 23/2323 port of a device for the purpose of determining whether the port is open. As can be seen from the Mirai source code, 160 IP addresses are randomly selected as target for the bot to scan. The bot sends TCP SYN packets to those IP addresses by using a raw socket. Then the raw socket is checked so as to obtain SYN+ACK packets.

As long as the bot receives a SYN+ACK packet, it establishes a TCP connection with the IP address in the non-blocking mode and records the connection in a table. After checking the raw socket and recording the connections, the bot checks the status of all TCP connections in the table. If a TCP connection fails to be built within 5 s, the bot will give up the connection request.

Conversely, if the TCP connection to the telnet port of an IP is established successfully, the bot will select a pair of usernames/password at random from the hardcoded list. Meanwhile, according to the feedback from the target device, the login can be verified as either a success

or a failure. If the bot receives no response from the device in 30 s, the login is deemed as a failure, and another trial goes on. The bot cuts off the unsuccessful connection and builds a new TCP connection to guess a random username/password again. However, such trials are only limited to ten times for Mirai to crack the username/password of a target IP.

Why Mirai Did Not Get Attention During Propagation

Since Mirai propagation did not cause the large-scale network congestion in the Internet in contrast to the propagation of Code Red worm (Zou et al. 2002), Mirai did not get attention during the propagation. To shed light on this phenomenon, we perform extensive simulations of Mirai propagation and analyze the time of TCP connection setup between a bot and a telnet port opened device to reflect the network state.

We simulate the original Mirai propagation using NS3. Recall that we divide the IP address space into three categories. We use empirical data to set the number of the devices in each category. Mirai grows to a peak of 600,000 infections at the end of November 2016 (Antonakakis et al. 2017), and the total number of IP addresses is 3,417,112,576 besides the IP addressed in the Mirai whitelists. Therefore, the proportion of vulnerable IoT devices is 0.0176%. Additionally, according to the results shown in ZoomEye (Zoo 2017), 11.1% online hosts open the telnet ports. To simplify the calculation, we round the proportions of the devices with open telnet ports and vulnerable telnet services to the nearest integer. Thus, in our simulation, we assume 10% of IoT devices in the IP space open the telnet port and 1% of the devices with the telnet service are vulnerable. Assume that the total number of devices in our simulated network is 65536. By using Monte Carlo method, we select 6325 devices that open telnet port, and 64 devices can be compromised by Mirai malware. Due to the limitation of computing power of IoT devices, the empirical results from Figure 6 in Antonakakis et al. (2017) show that the scanning bandwidth of 80% bots is smaller than 2000 Bps. Since the size of a TCP SYN scanning packet is 74 bytes, most bots can only probe around 30 IP addresses

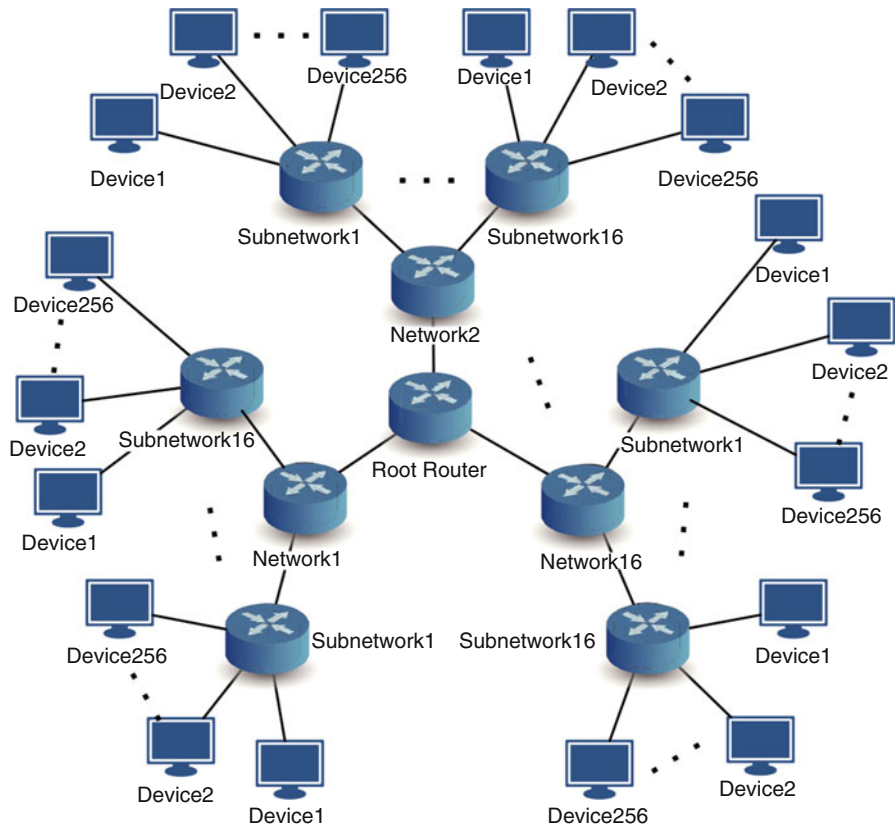
to scan their telnet ports in each round, rather than 160 IP addresses used in the Mirai source code. Thus, we change the scanning rate to 30 in our simulation. Figure 2 illustrates the network topology used in the simulation experiment. We use the NS3 global routing by building a global routing database for the topology using a Dijkstra Shortest Path First (SPF) algorithm, and we set the throughput of devices and routers as 1000 Mbps. The TCP connection setup time between a bot and a telnet port opened device is used to determine if the network is congested. Figure 3 illustrates that the Mirai does not cause network congestion by using empirical parameters in the simulation. Therefore, we can infer that the infected devices cannot generate a large amount of propagation traffic due to the limited computing power of vulnerable IoT devices.

Besides the limitation of IoT device computing power, another reason why Mirai does not cause network congestion is that the proportions of devices with open telnet port and vulnerable telnet service, respectively, are small. We assume the bot can probe 160 IP addresses in each round as the source code set and set two proportions in Table 1. The NS3 simulation results in Fig. 3 show that until we set 40% of IoT devices in the IP space with open telnet port and 50% of the devices with telnet service that are vulnerable, the time of TCP connection setup becomes longer. At that point, the Mirai propagation has caused network congestion. Therefore, Mirai cannot cause network congestion by scanning and infecting a small number of devices with open telnet port and vulnerable telnet service.

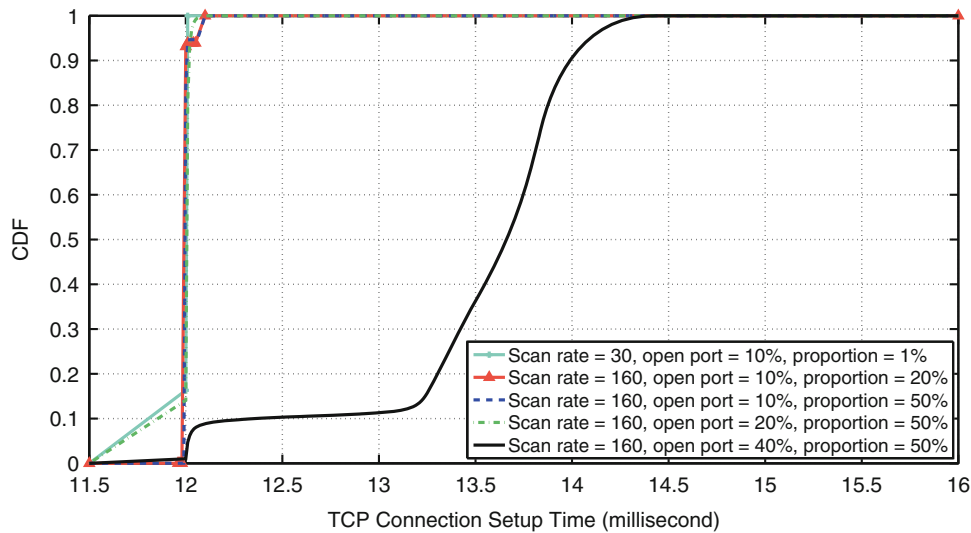
According to the real-world observation (Antonakakis et al. 2017), the proportion of devices with open telnet port and vulnerable telnet service are much lower than the 40% and 50%, respectively, and the limitation of IoT devices computing power cannot probe 160 IP addresses every round either. Therefore, Mirai did not cause network congestion and get attention during the propagation.

New Mirai Variants

We design new scanning strategies in order to boost the propagation. In contrast to the malware



New Variants of Mirai and Analysis, Fig. 2 Network topology used in the simulation experiment



New Variants of Mirai and Analysis, Fig. 3 TCP connection setup time distribution with different configurations

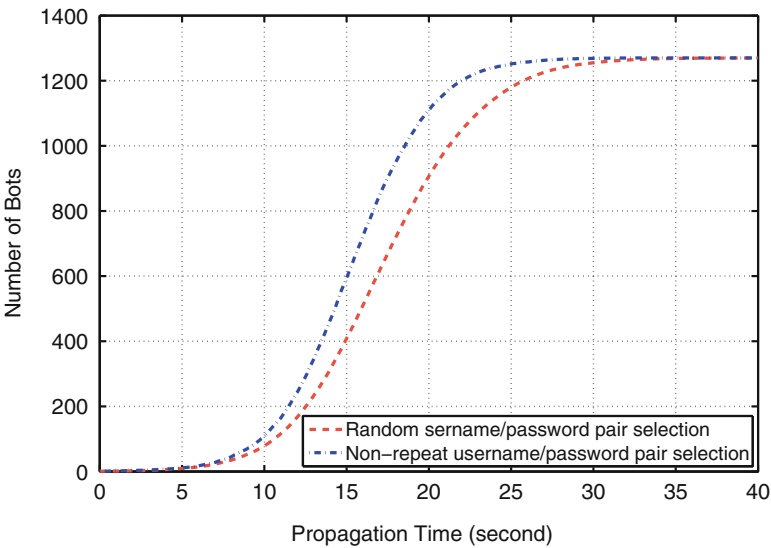
such as the Code Red worm (Zou et al. 2002) that exploits software vulnerabilities, the Mirai malware not only scan the devices to discover

the open telnet port but also crack the username/password using a hardcoded dictionary so as to successfully infect IoT devices. Therefore, the propagation speed of Mirai is much slower than that of the Code Red worm.

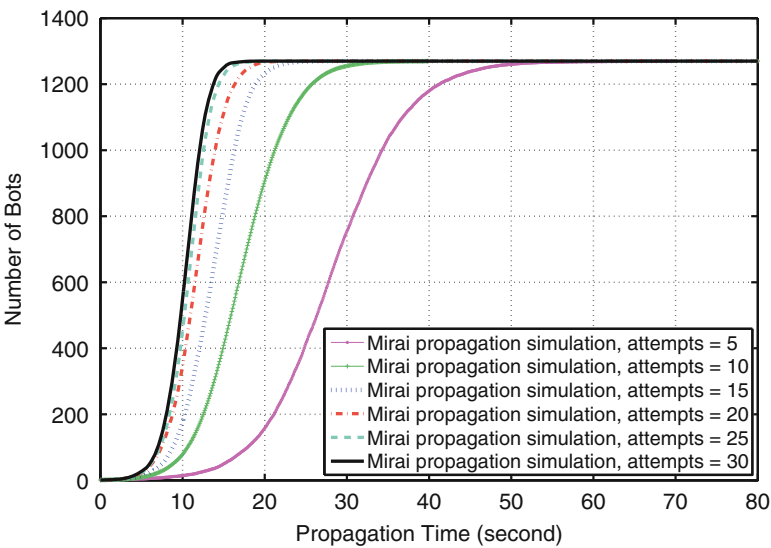
New Variants of Mirai and Analysis, Table 1 The proportions of devices with open telnet port and vulnerable telnet service

Devices with open telnet port	Devices with telnet service are vulnerable
10%	20%
10%	50%
20%	50%
40%	50%

New Variants of Mirai and Analysis, Fig. 4 The Mirai propagation speed using different cracking strategies



New Variants of Mirai and Analysis, Fig. 5 The Mirai propagation speed using different password cracking times



tries at most 10 times on a target, we can derive $q = [1 - (61/62)^{10}]$. Our variant of Mirai does not try the same password/username pair more than once. If the bot tries at most ten times on a target as the original Mirai, q can be improved to $q = 10/62$. The simulation result in Fig. 4 shows that non-repeat username/password pair try strategy can improve the Mirai propagation speed.

The value of q varies as the username/password cracking times change. Based on the analysis of the original Mirai scanning strategy, we carry out Mirai propagation simulations on a scale of username/password cracking times. The cracking times are set at 5, 15, 20, 25, and 30, respectively. The simulation results are displayed in Fig. 5. It is observed that with the number of attempts increased (i.e., q increases), the rise of the propagation speed ensues. Another observation is that as the number of attempts increases, the gain (acceleration) of the propagation speed falls. Thus we can devise a new variant of Mirai by keeping the username/password cracking times at 20 so as to fasten the Mirai propagation speed.

Future Directions

In this chapter, we study Mirai and propose new Mirai variants. We carefully introduce the propagation process of Mirai. Extensive NS3 simulations suggest that if the number of vulnerable devices is small, infected devices would not cause network congestion. This is why Mirai did not get attention during its propagation. To understand the potential impact of future Mirai attacks, we study new scanning strategies of Mirai that may adopt a non-repeating username/password cracking strategy and raise the number of cracking attempts so as to significantly boost the propagation speed. Our work in this chapter raises the alarm again for the IoT device manufacturers to better secure their products.

References

Anna-senpai (2016) [FREE] world's largest net:Mirai botnet, client, echo loader, CNC source code release. <https://hackforums.net/showthread.php?tid=5420472>

- Antonakakis M, April T, Bailey M, Bernhard M, Bursztein E, Cochran J, Durumeric Z, Halderman JA, Invernizzi L, Kallitsis M, Kumar D, Lever C, Ma Z, Mason J, Menscher D, Seaman C, Sullivan N, Thomas K, Zhou Y (2017) Understanding the Mirai Botnet. In: Proceedings of the 26th USENIX security symposium (Security)
- Department of Justice (2017) Justice department announces charges and guilty pleas in three computer crime cases involving significant DDoS attacks. <https://www.justice.gov/opa/pr/justice-department-announces-charges-and-guilty-pleas-three-computer-crime-cases-involving>
- Department of Justice (2018) Hackers' cooperation with FBI leads to substantial assistance in other complex cybercrime investigations. <https://www.justice.gov/usao-ak/pr/hackers-cooperation-fbi-leads-substantial-assistance-other-complex-cybercrime>
- Herzberg B, Bekerman D, Zeifman I (2016) Breaking down Mirai: an IoT DDoS botnet analysis. <https://www.incapsula.com/blog/malware-analysis-mirai-ddos-botnet.html>
- Hilton S (2016) Dyn analysis summary of Friday October 21 attack. <https://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack>
- Kambourakis G, Kolias C, Stavrou A (2017) The Mirai botnet and the IoT zombie armies. In: Proceedings of the IEEE military communications conference (Milcom)
- Klaba O (2016) Octave klaba twitter. <https://twitter.com/olesovhcom/status/778830571677978624>
- Krebs B (2016) KrebsOnSecurity hit with record DDoS. <https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos>
- Ling Z, Liu K, Xu Y, Gao C, Jin Y, Zou C, Fu X, Zhao W (2018) IoT security: an end-to-end view and case study. <https://arxiv.org/pdf/1805.05853.pdf>
- Mal (2016) Linux/Mirai, how an old ELF malware is recycled. <http://blog.malwaremustdie.org/2016/08/mmd-0056-2016-linuxmirai-just.html>
- Williams C (2016) Today the web was broken by countless hacked devices. https://www.theregister.co.uk/2016/10/21/dyn_dns_ddos_explained
- Zoo (2017) Zoomeye. <https://www.zoomeye.org/>
- Zou CC, Gong W, Towsley D (2002) Code red worm propagation modeling and analysis. In: Proceedings of the 9th ACM conference on computer and communication security (CCS)

Next-Generation Wireless Networks

► Big Data in 5G

NOMA

► Non-orthogonal Multiple Access

Non-orthogonal Multiple Access

Zekun Zhang¹, Haijian Sun¹, and Xianfu Lei²

¹Utah State University, Logan, UT, USA

²Southwest Jiaotong University, Chengdu, China

Synonyms

NOMA; PD-NOMA; Power domain NOMA

Definition

Non-Orthogonal multiple access (NOMA) is a multiple access method proposed by NTT DoCoMo for the next-generation (5G) mobile network. NOMA is a type of non-orthogonal multiplexing where users are multiplexed in the power domain, which is not sufficiently utilized by previous wireless mobile systems. Specifically, signals from different users are superposed at the transmitter side and separated at receiver by using successive interference cancellation (SIC). Note that in some articles, “NOMA” is used as a general name for any scheme that multiplexes users in a non-orthogonal manner, instead of the specific scheme described above. In such articles NOMA proposed by NTT DoCoMo normally is renamed as power domain NOMA (PD-NOMA).

Historical Background

Multiple access (MA) scheme has been regarded as a key technology to distinguish each generation of wireless communication systems in the past decades. In the existing wireless systems,

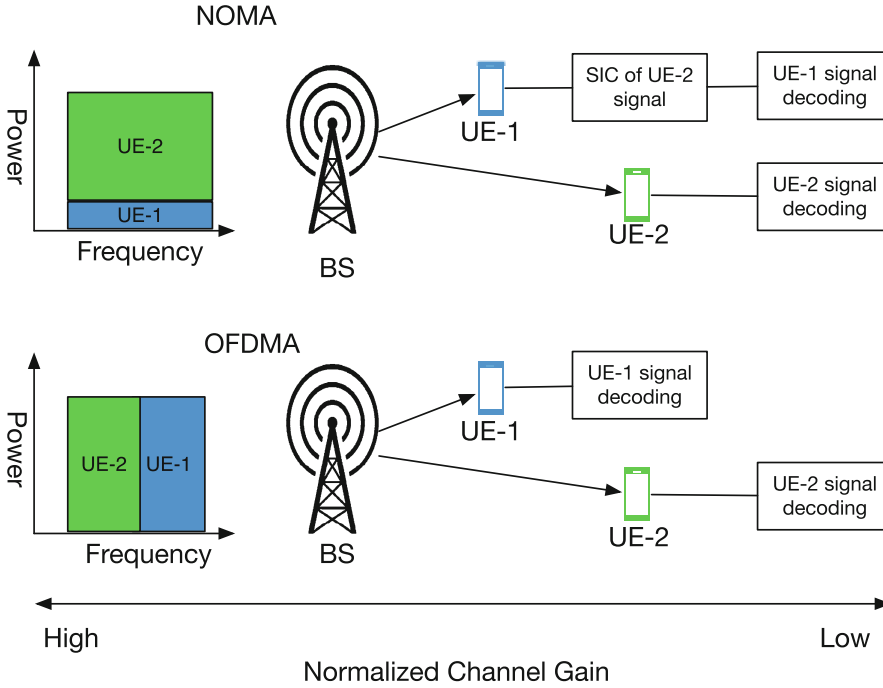
orthogonal multiple access (OMA) has been widely used from first-generation (1G) mobile communication system to 4G. Well-known frequency division multiple access (FDMA) for 1G, time division multiple access (TDMA) for 2G, code division multiple access (CDMA) for 3G, and orthogonal frequency division multiple access (OFDMA) for 4G are all primarily OMA schemes. In these OMA schemes, users are multiplexed orthogonally in either the frequency, time, or code domain so that mixed signal can be separated at receiver with low complexity.

While OMA can effectively minimize inter-user interference with a relatively low implementation complexity, its spectral efficiency needs to be further improved in order to meet the requirements in next-generation (5G) wireless mobile network such as supporting ultrahigh data throughput, massive connectivity, low latency, etc. To accommodate such demands, NTT DoCoMo proposed a downlink NOMA scheme (Higuchi and Kishiyama 2012; Saito et al. 2013; Benjebbour et al. 2015) where multiple users are multiplexed in the power domain at transmitter side and demultiplexed at receiver side by using SIC (Mazen et al. 2003).

The standardization of NOMA starts from being proposed to 3GPP LTE Release 13 (3GPP 2014). Then a study item (SI) under the name of “multiuser superposition transmission (MUST),” which is a downlink version of NOMA, is approved (3GPP 2015a). More details on MUST can be found in 3GPP (2015b). Though initially proposed as a downlink scheme, existing works show that NOMA can also be applied in uplink (Zhang et al. 2017; Takeda and Higuchi 2011), and the standardization of uplink NOMA is currently under evaluation (3GPP 2016).

Foundations

Figure 1 illustrates the downlink NOMA with a simple one base station (BS) and two user equipments (UEs) case. The transmission bandwidth allocated to two UEs is assumed to be 1 Hz. In NOMA, signals for two UEs are superposed by using Superposition Coding (SC) (Cover 1971)



Non-orthogonal Multiple Access, Fig. 1 Simple comparison between basic downlink NOMA and OMA (OFDMA)

at transmitter side. Denoting the signal for UE- i , $i \in \{1, 2\}$, as x_i where $\mathbb{E}[|x_i|^2] = 1$, the coded signal at transmitter is

$$x = \sqrt{P_1}x_1 + \sqrt{P_2}x_2.$$

P_i is the transmit power allocated to UE- i , and the sum of P_i is restricted to P , i.e., $P_1 + P_2 = P$. In this example P_i is set as $P_1 = 0.2P$ and $P_2 = 0.8P$. Then the received signal at UE- i is

$$y_i = h_i x + w_i,$$

where w_i denotes that Gaussian noise including intercell interference and the power density of w_i is N_0 . h_i is the complex channel coefficient from BS to UE- i . In downlink NOMA, SIC is implemented at UEs, and the optimal decoding order is in the order of increasing channel gain normalized by noise and intercell interference, i.e., $\frac{|h_i|^2}{N_0}$ (Saito et al. 2013; Benjebbour et al. 2015).

In the two-UE case above, assuming $\frac{|h_1|^2}{N_0} > \frac{|h_2|^2}{N_0}$, UE-2 does not need to perform SIC since it

comes first in the decoding order. Therefore, UE-2 decodes its received signal x_2 directly treating x_1 as interference, and the throughput of UE-2 can be represented as

$$R_2 = \log_2 \left(1 + \frac{P_2|h_2|^2}{P_1|h_2|^2 + N_0} \right).$$

Different with UE-2, UE-1 first decodes x_2 to subtract its component from the received signal y_1 . The throughput at UE-1 decoding x_2 can be then represented as

$$R_{1 \rightarrow 2} = \log_2 \left(1 + \frac{P_2|h_1|^2}{P_1|h_1|^2 + N_0} \right).$$

As $\frac{|h_1|^2}{N_0} > \frac{|h_2|^2}{N_0}$, $R_{1 \rightarrow 2} > R_2$ is guaranteed which means x_2 can be completely decoded at UE-1 assuming SIC process at UE-1 is error-free. After subtracting x_2 from y_1 , the throughput of UE-1 decoding its own signal x_1 is

$$R_1 = \log_2 \left(1 + \frac{P_1|h_1|^2}{N_0} \right).$$

In comparison with NOMA, the throughput of each UE in OMA (OFDMA) systems shown in Fig. 1 can be derived as

$$R_1^o = \frac{1}{2} \log_2 \left(1 + \frac{P|h_1|^2}{N_0} \right),$$

$$R_2^o = \frac{1}{2} \log_2 \left(1 + \frac{P|h_2|^2}{N_0} \right),$$

assuming the bandwidth and the transmit power are allocated to each UE equally. When $\frac{|h_1|^2}{N_0} = 20$ dB and $\frac{P|h_2|^2}{N_0} = 0$ dB, (Saito et al. 2013) numerically compared throughput of each UE between NOMA and OMA and demonstrated that the corresponding gains of NOMA from OMA are 32% and 48% for UE-1 and UE-2, respectively. Besides the numerical example shown here, the performance gain of NOMA is also theoretically analyzed in Ding et al. (2014).

The gain of NOMA comes from the multiplexing gain which is translated from the channel gain difference between two UEs through using SC at transmitter side and SIC at receiver side. Although in NOMA the transmit power allocated to a single UE can be lower than that in OMA, i.e., the transmit power of UE-1 is $0.2P$ and $0.5P$ for NOMA and OMA, respectively, both UEs can benefit from being scheduled more bandwidth. While NOMA relies on advanced receiver processing ability such as SIC, the expectation for evolution of processing ability of user devices is reasonable, generally following Moore's law. The merits of NOMA can be summarized as following (Dai 2015; NTT-DOCOMO 2014):

1. Improved spectral efficiency: As shown in the example above, NOMA can offer higher spectral efficiency for both UEs. This advantage of NOMA over OMA can also be explained from the perspective of information theory. As proofed in Tse and Viswanath (2005), NOMA with SC at transmitter side and SIC at receiver side can achieve the optimal capacity region of the downlink broadcast channel. However, OMA is not able to achieve so.
2. Massive connectivity: Different with OMA, the number of supported users in NOMA is

not strictly limited by the amount of available resources and their scheduling granularity. Therefore, NOMA can be used to address the challenges of massive connectivity.

3. Robust performance gain: NOMA transmitter does not rely much on instantaneous channel state information (CSI), which requires feedback signaling from UEs, to perform multiplexing. Therefore the robust performance gain can be expected irrespective UE mobility and signaling latency.

Key Applications

There have been many standardization activities on the implementation of NOMA in 5G wireless mobile networks. Particularly, in 3GPP (2015b) the performance of MUST, the downlink version of NOMA, is comprehensively studied, and the conclusions are drawn as:

- MUST can increase system capacity as well as improve user experience in certain scenarios.
- MUST is generally more beneficial when the network experiences higher traffic load.
- MUST is generally more beneficial in user-perceived throughput for wideband scheduling case, compared to subband scheduling case.
- MUST is generally more beneficial in user-perceived throughput for cell-edge UEs, compared to other UEs.
- MUST-far UEs can be legacy UEs when QPSK is applied to MUST-far UEs or the most two significant bits in the modulation symbol are assigned to far UE.

Based on the advantages of MUST listed above, the implementation of MUST in some scenarios of 5G networks can be envisaged such as machine-to-machine (M2M) communications, ultradense networks (UDN), and mMTC (Recommendation ITU-R M.2083: IMT Vision 2015; Ding et al. 2017) where massive connections and the IoT functionality of 5G are required.

Cross-References

- ▶ [Cognitive Radio-Based Non-orthogonal Multiple Access](#)
- ▶ [Densification of Wireless Networks, from Few to Massive](#)
- ▶ [Millimeter Wave NOMA](#)
- ▶ [Multiple Access Technique for Cellular Wireless Networks](#)

References

- 3GPP (2014) Justification for NOMA in new study on enhanced multi-user transmission and network assisted interference cancellation for LTE. RP-141936, Dec 2014
- 3GPP (2015a) New SI proposal: study on downlink multiuser superposition transmission for LTE. RP150496, Mar 2015
- 3GPP (2015b) Study on downlink multiuser superposition transmission (MUST) for LTE (release 13), TR36.859, Dec 2015
- 3GPP, NTT-DOCOMO (2016) Initial views and evaluation result on non-orthogonal multiple access for NR uplink. R1-163111, Apr 2016
- Benjebbour A, Li A, Saito K et al (2015) NOMA: from concept to standardization. Paper presented at IEEE conference on standards for communications and networking, University of Tokyo, Tokyo, 28–30 Oct 2015
- Cover T (1971) Broadcast channels. *IEEE Trans Inf Theory* 18(1):2–14
- Dai L, Wang B, Yuan Y et al (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun Mag* 53(9):74–81
- Ding Z, Yang Z, Fan P et al (2014) On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Signal Process Lett* 21(12):1501–1505
- Ding Z, Lei X, Karagiannidis GK et al (2017) A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. *IEEE J Sel Areas Commun*. <https://doi.org/10.1109/jsac.2017.2725519>
- Higuchi K, Kishiyama Y (2012) Non-orthogonal access with successive interference cancellation for future radio access. Paper presented at the 9th IEEE vehicular technology society Asia pacific wireless communications symposium, Kyoto University, Kyoto, 23–24 Aug 2012
- Mazen OH, Alouini M-S, Bastami A et al (2003) Performance analysis of mobile cellular systems with successive co-channel interference cancellation. *IEEE Trans Wireless Commun* 2(1):29–40
- NTT-DOCOMO (2014) DOCOMO 5G white paper. https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_5g/DOCOMO_5G_White_Paper.pdf. Accessed July 2014
- Recommendation ITU-R M.2083: IMT Vision (2015) Framework and overall objectives of the future development of IMT for 2020 and beyond, Sept 2015
- Saito Y, Kishiyama Y, Benjebbour A et al (2013) Non-orthogonal multiple access (NOMA) for cellular future radio access. Paper presented at IEEE 77th vehicular technology conference: VTC 2013-Spring, Dresden, 2–5 June 2013
- Takeda T, Higuchi K (2011) Enhanced user fairness using non-orthogonal access with SIC in cellular uplink. Paper presented at IEEE 74th vehicular technology conference: VTC2011-Fall, San Francisco, 5–8 Sept 2011
- Tse D, Viswanath P (2005) Fundamentals of wireless communication. Cambridge University Press, Cambridge
- Zhang Z, Sun H, Hu H (2017) Downlink and uplink non-orthogonal multiple access in a dense wireless network. *IEEE J Sel Areas Commun*. <https://doi.org/10.1109/jsac.2017.2724646>

Non-orthogonal Multiple Access (NOMA)

Yong Zhou¹ and Vincent W. S. Wong²

¹School of Information Science and Technology, ShanghaiTech University, Shanghai, China

²Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada

Synonyms

[Successive interference cancellation](#); [Superposition coding](#); [User pairing](#)

Definitions

Non-orthogonal multiple access (NOMA) refers to multiple access schemes that can enable two or more users to be served by the same base station in a single orthogonal resource block (e.g., time slot, subcarrier, spreading code) simultaneously.

Background

Due to the rapidly increasing traffic demand and the scarcity of the radio spectrum, multiple access is necessary for cellular systems to serve a large number of users. Each generation of wireless cellular systems is characterized by a specific multiple access scheme. In particular, frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA) are utilized in the first-generation (1G), the second-generation (2G), the third-generation (3G), and the fourth-generation (4G) systems, respectively. As FDMA, TDMA, and OFDMA are orthogonal multiple access (OMA) schemes, each orthogonal resource block in the frequency and time domains can only be used to serve a single user, which is not spectrally efficient. On the other hand, CDMA requires that the chip rate has to be much higher than the supported data rate, which is very difficult to be implemented when the supported data rate is as high as 10 Gbps. These multiple access schemes are not able to support the typical use cases of the fifth-generation (5G) systems, which have very diverse quality of service (QoS) requirements, e.g., extremely high data rates, massive connection density, and ultralow latency.

As a promising multiple access technique to meet the QoS requirements of 5G systems, non-orthogonal multiple access (NOMA) has recently received considerable attention (Saito et al. 2013; Ding et al. 2014; Wong et al. 2017). The key idea of NOMA is to allow two or more users to be served in one orthogonal resource block simultaneously so as to enhance the spectral efficiency. Different variants of NOMA schemes have been proposed, including power-domain NOMA (Saito et al. 2013), sparse code multiple access (SCMA) (Nikopour and Baligh 2013), pattern division multiple access (PDMA), and low density spreading (LDS). Power domain NOMA is a single-carrier NOMA scheme, where the NOMA users' signals are transmitted in the same subcarrier. On the other hand, SCMA, PDMA, and LDS are multi-carrier NOMA

schemes, where each NOMA user's signal is spread over multiple subcarriers. Due to its simplicity and efficiency, power-domain NOMA is discussed in the rest of this entry.

Power-Domain NOMA

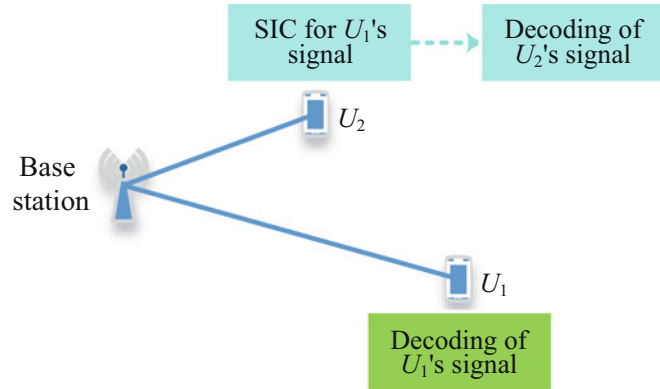
With power-domain NOMA, multiple users can be served simultaneously by a single base station using the same frequency channel and spreading code by allocating different transmit powers to the users. Power-domain NOMA can be applied in both downlink and uplink transmissions. Consider the two-user downlink NOMA scheme as an example, as shown in Fig. 1. The base station superimposes the signals intended for two users. The user experiencing a worse channel condition (e.g., U_1) is allocated a higher transmit power than the user experiencing a better channel condition (e.g., U_2). The user allocated a higher transmit power (e.g., U_1) decodes its own signal directly by treating the signal intended for the other user (e.g., U_2) as co-channel interference. On the other hand, the user allocated a lower transmit power (e.g., U_2) decodes its own signal after performing successive interference cancellation (SIC) to remove the other user's signal.

By accommodating two or more users in one orthogonal resource block, NOMA has the potential to significantly enhance the spectral efficiency, which can be utilized to achieve high system throughput (Zhou and Wong 2017), to support massive connectivity (Mostafa et al. 2017), and to reduce the medium access delay (Liang et al. 2017). Compared to OMA, NOMA can achieve a balance between system throughput and user fairness. Due to these advantages, NOMA, termed as multiuser superposition transmission (MUST), has been specified in the 3rd Generation Partnership Project (3GPP) Long-Term Evolution Advanced (LTE-A) standard (3GPP 2016). In addition, NOMA, termed as layered division multiplexing (LDM), has also been included in the digital TV standard (ATSC 3.0) (Zhang et al. 2016).

Transmit power allocation among the paired NOMA users plays an important role in determin-

Non-orthogonal Multiple Access (NOMA),

Fig. 1 Downlink power-domain NOMA transmission



ing the performance of power-domain NOMA. The outage probability and the sum rate of power-domain NOMA under the fixed and cognitive radio based transmit power allocation schemes were studied in Ding et al. (2016). An optimal transmit power allocation scheme was proposed in Choi (2016) to maximize the sum rate of the paired NOMA users when either the instantaneous or statistical channel state information (CSI) is available at the base station. On the other hand, user pairing is another aspect that is critical to the performance of NOMA. It has been demonstrated in Ding et al. (2016) that the performance gain of NOMA over OMA is greater when the users with more diverse channel conditions are paired. Based on the level of CSI (e.g., instantaneous channel gain, user's distance) available at the base station, different user pairing strategies can be utilized.

Cooperative NOMA and mmWave-NOMA

Due to its compatibility, NOMA can be combined with many communication technologies, including cooperative communication, millimeter-wave (mmWave), multiple-input multiple-output (MIMO), and visible light communication (VLC), to further enhance the network performance. For simplicity of explanation, two-user power-domain NOMA scheme is considered here. For two-user NOMA, the user located closer to the base station is

referred to as the near user, and the other user is referred to as the far user. Both cooperative NOMA and mmWave-NOMA schemes are briefly discussed as follows.

1. **Cooperative NOMA:** Based on the principle of NOMA, the far user shares the radio channel and the transmit power with the near user. Compared to OMA, the achievable performance of the far user is degraded. As the near user has to decode the signal intended for the far user before decoding its own signal, the near user can act as a relay and help forward the signal to the far user, which can effectively enhance the reception reliability of the far user.

By exploiting this feature, a cooperative NOMA scheme was proposed in Ding et al. (2015). The cooperative NOMA transmission can be divided into two phases. In the first phase, the base station broadcasts the superimposed mixture of the signals. In the second phase, the near user acts as a half-duplex relay and forwards the signal intended for the far user. After receiving the signal, the far user can decode its own signal by utilizing different combining techniques (e.g., selection combining, maximum ratio combining). The half-duplex cooperative NOMA scheme requires an additional time slot for relay transmission. To address this issue, a dynamic decode-and-forward (DDF)-based cooperative NOMA scheme was proposed in Zhou et al. (2017), which allows a half-duplex relay to provide a cooperative diversity gain without

the need of additional time slots. In particular, the near user decodes the superimposed signals based on partial reception, where the reception duration depends on the instantaneous channel gain. As soon as both signals are successfully received, the near user switches to the transmit mode and forwards the signal to the far user until the end of the time slot. As a result, both the superpositioned transmission and the cooperative diversity gain can be achieved in one time slot.

2. **mmWave-NOMA:** Due to the large amount of spectrum resources available in mmWave frequency bands, mmWave communication is a key enabling technology for the 5G systems. It is suitable for small-cell communications with coverage area of 200 m. Different from sub-6 GHz frequencies, mmWave frequencies (i.e., 30–300 GHz) suffer from higher path loss and penetration loss. Due to a smaller wavelength, multiple antennas can be deployed at the base station to achieve directional beamforming and exploit the array gain to ensure sufficient received signal power. The coexistence of mmWave and NOMA techniques for the 5G systems has recently been studied (Ding et al. 2017a,b).

Due to the directivity of mmWave transmission, the base station has to align its beam with the paired NOMA users' channels for the best system performance, which would introduce non-negligible pilot overhead. In order to reduce the protocol overhead required for channel estimation, the authors in Ding et al. (2017b) proposed an mmWave-NOMA scheme with random beamforming. Specifically, the base station equipped with multiple antennas forms transmit beams based on a random angle. Tools from stochastic geometry are utilized to analyze the outage probability and sum rate. By taking into account the hardware constraint that a practical circuit can support a limited number of phase shifts, the authors in Ding et al. (2017a) proposed a finite resolution analog beamforming scheme in massive MIMO and mmWave networks. Although the transmit beams are not perfectly aligned with the users' channels, mmWave-

NOMA can still achieve higher sum rates of the paired NOMA users than mmWave-OMA.

Conclusions

NOMA is a key enabling technique for 5G systems, and it can be combined with many communication techniques (e.g., massive MIMO, cooperative communication, mmWave) to achieve high system throughput, massive connectivity, and low latency. This entry reviewed the development, the basic concept, and the applications of NOMA. The examples of cooperative NOMA and mmWave-NOMA demonstrated the performance gain of NOMA over OMA.

Cross-References

- [Capacity Analysis of Interference-Limited Wireless Networks in Three-Dimensional Space](#)
- [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)
- [Resource Allocation in NOMA-Assisted D2D Networks](#)

References

- 3GPP (2016) Study on downlink multiuser superposition transmission (MUST) for LTE
- Choi J (2016) On the power allocation for MIMO-NOMA systems with layered transmissions. *IEEE Trans Wirel Commun* 15(5):3226–3237
- Ding Z, Yang Z, Fan P, Poor HV (2014) On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Signal Process Lett* 21(12):1501–1505
- Ding Z, Peng M, Poor HV (2015) Cooperative non-orthogonal multiple access in 5G systems. *IEEE Commun Lett* 19(8):1462–1465
- Ding Z, Fan P, Poor HV (2016) Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. *IEEE Trans Veh Technol* 65(8):6010–6023
- Ding Z, Dai L, Schober R, Poor HV (2017a) NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks. *IEEE Commun Lett* 21(8):1879–1882
- Ding Z, Fan P, Poor HV (2017b) Random beamforming in millimeter-wave NOMA networks. *IEEE Access* 5:7667–7681

- Liang Y, Li X, Zhang J, Ding Z (2017) Non-orthogonal random access for 5G networks. *IEEE Trans Wirel Commun* 16(7):4817–4831
- Mostafa AE, Zhou Y, Wong VWS (2017) Connectivity maximization for narrowband IoT systems with NOMA. In: *Proceeding of IEEE ICC, Paris, May 2017*
- Nikopour H, Baligh H (2013) Sparse code multiple access. In: *Proceeding of IEEE PIMRC, London, Sep 2013*
- Saito Y, Kishiyama Y, Benjebbour A, Nakamura T, Li A, Higuchi K (2013) Non-orthogonal multiple access (NOMA) for cellular future radio access. In: *Proceeding of IEEE VTC Spring, Dresden, Jun 2013*
- Wong VWS, Schober R, Ng DWK, Wang LC (2017) *Key Technologies for 5G Wireless Systems*. Cambridge University Press, West Nyack
- Zhang L, Li W, Wu Y, Wang X, Park SI, Kim HM, Lee JY, Angueira P, Montalban J (2016) Layered-division-multiplexing: theory and practice. *IEEE Trans Broadcast* 62(1):216–232
- Zhou Y, Wong VWS (2017) Stable throughput region of downlink NOMA transmissions with limited CSI. In: *Proceeding of IEEE ICC, Paris, May 2017*
- Zhou Y, Wong VWS, Schober R (2017) Performance analysis of cooperative NOMA with dynamic decode-and-forward relaying. In: *Proceeding of IEEE Globecom, Singapore, Dec 2017*

OFDM

- [Index Modulation](#)

OFDM Synchronization

- [Time and Frequency Synchronization in OFDM Systems](#)

OFDM/OFDMA System Without CP

- [Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation](#)

Offloading Delay-Tolerable Data Traffic to Connected Vehicle Networks

Pengbo Si and Yanhua Zhang
Faculty of Information Technology, Beijing
University of Technology, Beijing, China

Synonyms

[Vehicular network data offloading](#)

Definitions

A new architecture is introduced to efficiently utilize the potential resource from connected vehicles and to mitigate the congestion problem in other data networks. Delay-tolerable data traffic is offloaded from the data networks to the connected vehicle networks without extra infrastructure/hardware deployment. An optimal distributed data hopping mechanism is also introduced to enable delay-tolerable data routing over connected vehicle networks. The next-hop decision optimization problem is formulated as a partially observable Markov decision process (POMDP). Simulation results are also presented to demonstrate the performance improvement of the proposed scheme.

Historical Background

Boosted by the rapid growth of data-generating devices, such as various types of sensors, smartphones, tablets, video game consoles, surveillance cameras, augmented reality devices, and wearable electronics, among others, the exponential increase in the volume of data has imposed significant pressure on the current communication networks (Szymanski 2013). Congestions at backhaul, backbone, and access networks could cause serious degradation in performances including data rate, packet loss rate, and delay, especially for the services with strict quality-of-service (QoS) and/or quality-of-

experience (QoE) requirements (Habachi et al. 2013).

As resources become the key constraint in communication networks, recent efforts on improving the QoS/QoE are made in two directions: utilizing more resources and utilizing resources more efficiently (Zhang et al. 2014a). Infrastructure deployment provides extra bandwidth, processing, and storage resources for the networks, while millimeter wave and terahertz communication technologies are exploiting new available spectrum resource. Besides, promising technologies, such as small cell, cognitive radio, massive multiple-input multiple-output (MIMO), and network function virtualization (NFV), contribute significantly to improving the resource utilization efficiency. Nevertheless, higher service quality is achieved at the increased costs on capital expenditures (CAPEX) and operating expenses (OPEX), due to the deployment and maintenance of extra infrastructure or hardware introduced by the above technologies.

However, a large portion of data traffic does not require high QoS, and thus extra costs on QoS improvement for such traffic could be a waste (Koo et al. 2007). For example, in content delivery networks (CDNs), delay of hours is acceptable when the popular contents are pushed to the servers at the edge of the network. In big data industries, high delay would not seriously affect the performance of some data backup services. Another scenario is the personal data that is delivered from one local data center to another as the data owner travels from one location to another to minimize the data obtaining delay. This personal data delivery could possibly tolerate high delay as well. Note that such delay-tolerable traffic also has the delay constraints or lifetime. The only difference is that its tolerable delay is much higher than delay-sensitive traffic. Aiming at satisfying various QoS requirements of different types of services while minimizing the costs, priority-based schemes have been widely studied to deliver delay-sensitive traffic with higher priority and treat delay-tolerable traffic as best-effort services (Zhang et al. 2014b). Unfortunately, since the total amount of com-

munication resources is limited and the traffic load becomes heavy, delay-tolerable traffic may have very little delivery chance, and thus the probability of successful delivery within its delay constraints could be low.

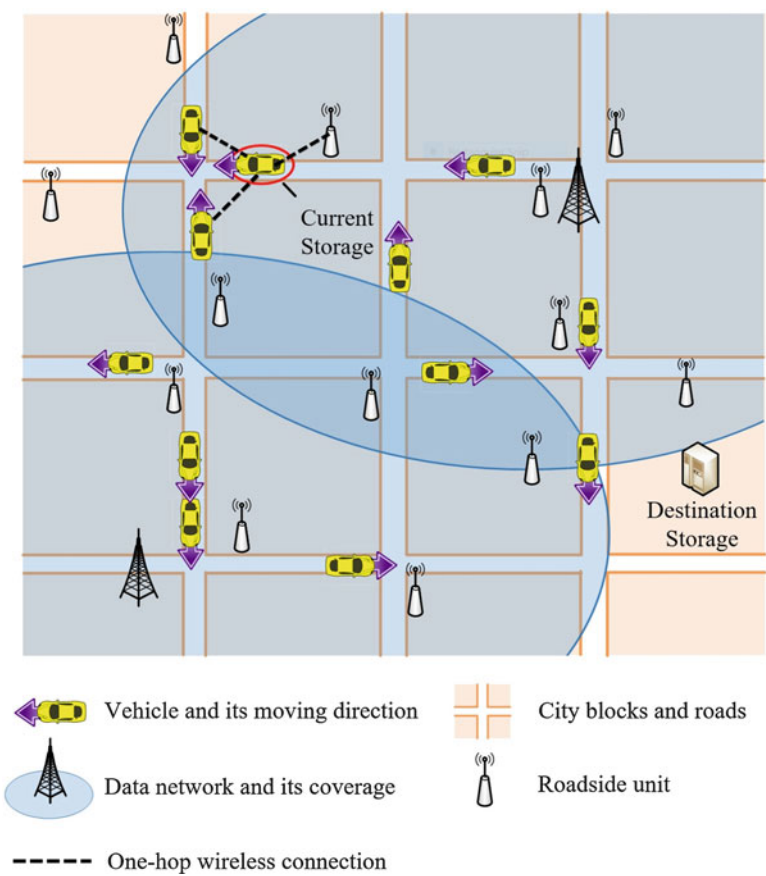
On the other hand, with the rapid development of technologies on connected vehicles and vehicular networks, millions of roadside units (RSUs) and vehicles equipped with communication, computing, storage, and positioning devices will be connected to form a worldwide network (Uhlemann 2015). Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications, enabled by IEEE 802.11p and dedicated short-range communication (DSRC) technologies, allow road safety, driver assistance, and infotainment data transmission with up to 27 Mbps data rate (Kenney 2011). Furthermore, vehicular delay-tolerant networks (DTNs) have been widely studied recently, considering the vehicle networks that may lack continuous network connectivity due to the highly dynamic topology. With the embedded storage and communication capabilities of the vehicles and RSUs, data can be delivered in a “store-carry-forward” manner, taking advantage of the mobility of the vehicles (Isento et al. 2013). An in-depth analysis of the different proposals for vehicular DTNs is studied, and the DTN proposals are classified according to the type of knowledge needed to route messages in Tornell et al. (2015). However, current research on vehicular DTNs ignores the interaction with other existing data networks and focuses on delivering the data generated from (such as road safety and driver assistance data) or required by (such as traffic condition and infotainment data) the vehicular networks themselves, of which the data traffic is light, and thus the vehicle network resource utilization efficiency is low.

Connected Vehicle Network for Delay-Tolerable Data Delivery

To efficiently utilize the potential resource from connected vehicles and to mitigate the congestion problem in data networks, a new architecture is

**Offloading
Delay-Tolerable Data
Traffic to Connected
Vehicle Networks, Fig. 1**

An example of the
proposed architecture



proposed, in which delay-tolerable data traffic is offloaded from the data networks to the connected vehicle networks without extra infrastructure/hardware deployment. Delay-tolerable data is optimally scheduled to deliver over data networks and/or connected vehicle networks, according to the delay constraints.

The architecture proposed in this chapter is presented in Fig. 1. Connected vehicles and RSUs with embedded computing, storage, and positioning devices cooperate with each other and form a connected vehicle network. IEEE 802.11p- and DSRC-based technologies are adopted for wireless communications among the vehicles and RSUs. In this network, road safety, driver assistance, and other vehicle-related information are transmitted with higher priority, while spare network resources are utilized to deliver delay-tolerable data. Besides, vehicles and RSUs are also connected to the Internet via

other data networks, including cellular, WiFi, asymmetric digital subscriber line (ADSL), and fiber-to-the-home (FTTH) networks, among others. Data networks usually have better QoS performance with higher costs, compared to the hop-by-hop communications in the connected vehicle network.

With this architecture, a delay-tolerable file, no matter it is from/to the connected vehicle network or not, is segmented into equal-size data blocks and delivered over the connected vehicle network and/or other data networks, depending on the network condition and the delay constraints of the data blocks. In other words, the connected vehicle network and other data networks cooperate to schedule the data block delivery, in such a manner that delay-sensitive data traffic is delivered roughly over data networks, and connected vehicle networks are utilized mainly for delay-tolerable data blocks.

In the connected vehicle network, each vehicle or RSU has the information on its location and moving direction and keeps the direct connection with its neighbors within its one-hop communication range. A delay-tolerable data block is delivered toward its destination in a hop-by-hop and distributed manner. At the beginning of each time slot, the current storage of the data block, which denotes the device that has the data block in its storage, makes the offloading and/or hopping decision on which network should be utilized to deliver the data block, whether to keep the data block in the current storage or not and, if not, which neighbor it should be transmitted to. The decisions are made according to the delay constraint, the location of the destination, and the moving information of the current storage and its neighbors. Then, the data block is kept in the current storage or transmitted to the next-hop device. If the deadline of a data block is close, data networks may be selected as the next hop to guarantee that the data block is delivered to its destination on time.

Problem Formulation

In this section, the next-hop device decision problem is formulated as a partially observable Markov decision process.

System State Space

Decisions are made according to the information on the current storage and the neighbors, as well as the distance and angle to the destination. Thus, we define the system state as a group of destination angle, the moving direction of the current storage, the moving indicator of the current storage, the distance to the destination $d(k)$, and the successful-transmission indicator of the neighbors. Here, the moving indicator of the current storage is set to 1 or 0, if the current storage is moving or not, respectively.

Actions and Policies

The action $a(k)$ of the current storage is the decision on the offloading and/or hopping. The current storage needs to decide whether the data block will be transmitted over the vehicle network or the data networks. If data networks are selected, the data block will be transmitted to the data network and then delivered to its destination in the next time slot. If not, the current storage makes the decision on whether to keep the data block in the current storage, and if not, which neighbor the data block should be transmitted to. Let $a(k = 1)$ denote that the data block is transmitted over data networks, $a(k) = 0$ denote that the data block is kept in the current storage, and $a(k) = \beta$ denote that the data block is transmitted to a neighbor with the moving direction β .

Observations and State Transitions

The successful-transmission indicator of the neighbors may not be fully observable at each time slot. Fortunately, it can be estimated according to the observations on the neighbors of the current storage. As the system state transits from $s(k)$ to $s(k + 1)$ under action $a(k)$, an observation $o(k)$ is generated with the conditional probability $\text{Prob.}(o(k)|s(k + 1), a(k))$. Under action $a(k)$, the system state $s(k)$ is a Markov process, with the one-step transition probability $\text{Prob.}(s(k + 1)|s(k), a(k))$, or $P_s^a(i, j)$ for short, where $\hat{s}_i$ and $\hat{s}_j$ are the realizations of $s(k)$, and $\hat{a}_i$ is the action taken at t_k .

Objective and Reward

Our objective is to minimize the costs and successfully deliver the data blocks to their destinations within the required time constraints. We use $R(k)$ to represent the immediate reward obtained in each time slot:

$$R(k) = \begin{cases} R_{\text{arrive}} - C_{\text{Veh}}, & \text{if } d(k) \neq 0, a(k) \neq 0, d(k + 1) = 0, \\ -C_{\text{Veh}}, & \text{if } d(k) \neq 0, a(k) \neq 0, d(k + 1) \neq 0, \\ R_{\text{arrive}} - C_{\text{DN}}(d(k)), & \text{if } d(k) \neq 0, a(k) = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where R_{arrive} is the reward obtained by successfully delivering the data block to its destination, C_{veh} is the cost to use the connected vehicle network for delivering the data block in each time slot, and $C_{\text{DN}}(d(k))$ is the cost to use the data networks for delivering the data block to the destination. Note that $C_{\text{DN}}(d(k))$ depends on the distance $d(k)$ from its current location to its destination.

Then, the decision optimization problem can be represented as:

$$g^* = \operatorname{argmax}_{g \in \mathbf{G}} \sum_{k=1}^K \gamma^{K-k} R(k), \quad (2)$$

where γ is the discount factor, $0 < \gamma < 1$. Finite horizon is considered for the data block transmissions. Consequently, no reward can be obtained if a data block fails to arrive at the destination before t_K , i.e., $\sum_{k=1}^K \gamma^{K-k} R(k) \leq 0$ if $d(K) \neq 0$.

Solving the POMDP Problem

With the system state space $\mathbf{S}$, action space $\mathbf{A}$, observation space $\mathbf{O}$, transition probabilities P , and system reward R , a POMDP tuple $\langle \mathbf{S}, \mathbf{A}, \mathbf{O}, P, R \rangle$ is formed. Many existing algorithms, including low-complexity heuristic ones, have been proposed to solve it by value and policy iterations and belief state updating and successfully applied in various scenarios, such as robotic control, task assignment, and radio resource management, among others (Padmanabhan et al. 2015; Young et al. 2013; Krishnamurthy 2011; Zhao et al. 2007; Si et al. 2009). Due to the space limit, details of these algorithms are omitted in this chapter.

Data Hopping Process

With the multitasking capability of the storage devices and the media access control (MAC) protocols that enable multiple orthogonal channels for sending and receiving, it is straightforward to assume that a storage device is capable to send and receive data blocks simultaneously. At the beginning of each time slot, the current storage device updates its observation and calculates the

optimal decision $a^*(k)$ based on the observations. Based on $a^*(k)$, the data block is kept in the local storage or transmitted to the next-hop device.

Simulation Results and Analysis

In this section, simulation results are presented to demonstrate the significant performance improvement of the proposed scheme.

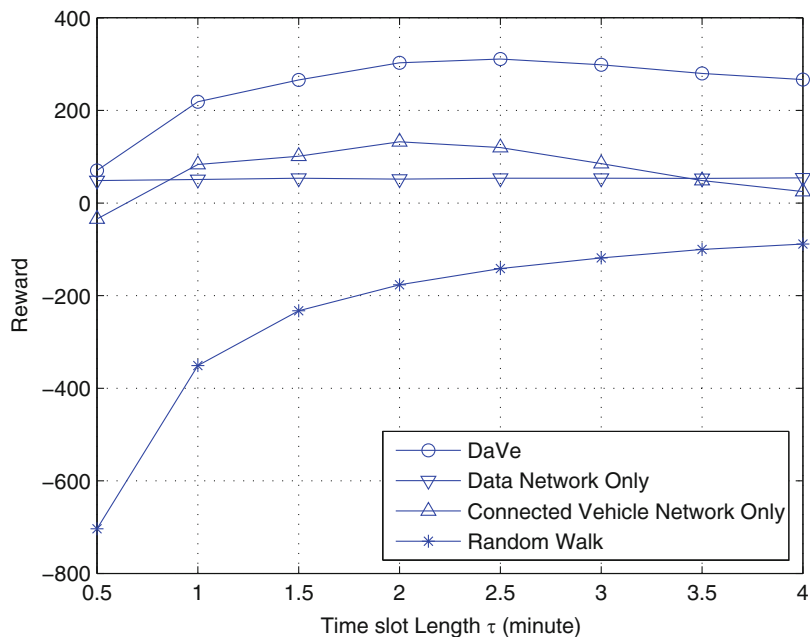
Simulation Scenario

In our simulations, we consider a $10,000 \times 10,000$ m area, where 20,000 vehicles move along the roads at the average velocity $v = 8.33$ m/s (30 km/h). The average distance between two adjacent intersections is 300 m. When a vehicle arrives at an intersection, it stops and waits with the probability $P_{\text{wait}} = 0.75$ and the maximum waiting time $T_{\text{wait}} = 120$ s. We assume that the delivery delay requirement of the delay-tolerable data blocks is 18,000 s (5 h) and the data block size $B = 480$ Mbits = 60 MBytes. DSRC technology is adopted to enable wireless communications between connected vehicles and/or RSUs, of which the communication range $L_{\text{range}} = 200$ m and the achievable data rate $H = 16$ Mbps. Besides, we adopt $\tau = 120$ s as the time slot length, $R_{\text{arrive}} = 500$ as the successful delivery reward, $C_{\text{veh}} = 1$ as the cost to use the connected vehicle network in each time slot, and $C_{\text{DN}}(d(k)) = 100d(k)$ as the cost to use the data network to deliver the delay-tolerable data block.

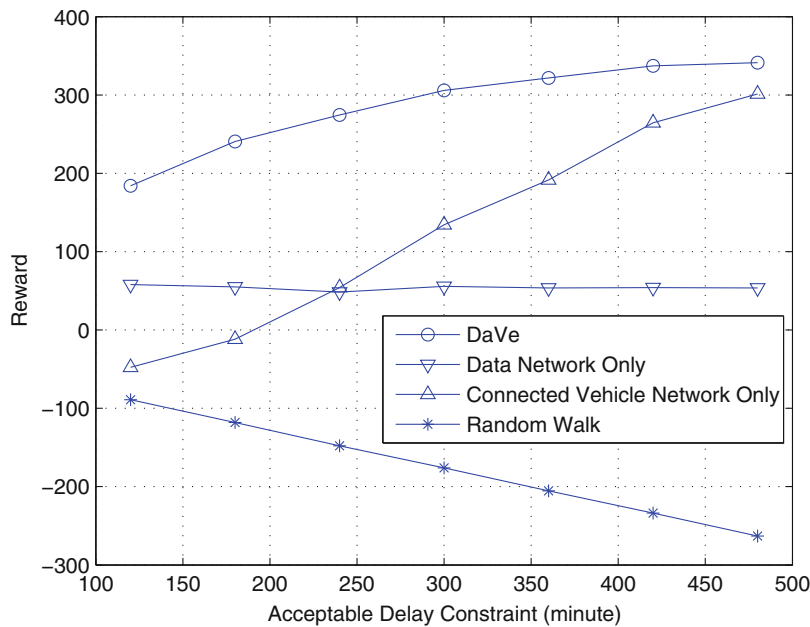
The performance of our proposed optimal scheme is compared with three other schemes, including “data network only” (only data network is utilized to deliver the data), “connected vehicle network only” (only connected vehicle network is utilized, with optimal data block scheduling), and “random walk” (only connected vehicle network is utilized, with random data block scheduling).

System Reward Comparison

In Figs. 2, 3, and 4, we compare the performance of system reward, which is defined by Eq. (1). With different time slot length from 0.5 to 4 min, Fig. 2 demonstrates that the proposed scheme



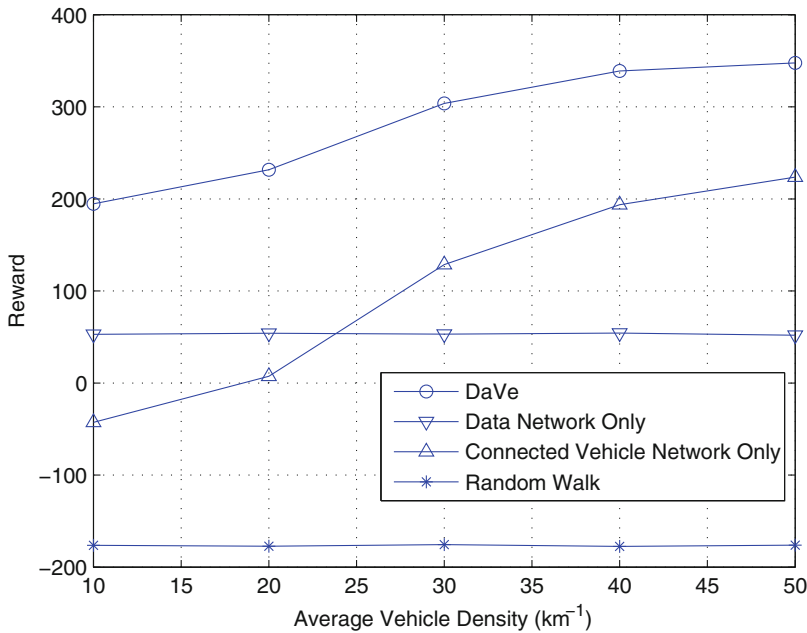
Offloading Delay-Tolerable Data Traffic to Connected Vehicle Networks, Fig. 2 System reward comparison with different time slot length τ



Offloading Delay-Tolerable Data Traffic to Connected Vehicle Networks, Fig. 3 System reward comparison with different data block delay requirement

provides the highest reward compared to the other three schemes. If only the connected vehicle network is used to deliver the delay-tolerable traffic,

the system reward is lower due to that some data blocks may not arrive at the destination storage within the required time limit and then R_{arrive} is



Offloading Delay-Tolerable Data Traffic to Connected Vehicle Networks, Fig. 4 System reward comparison with different vehicle density w_{veh}

not obtained. The reward of the “data network only” scheme does not change with the different values of time slot length. Data networks always guarantee the short delay of the data blocks, and the costs depend only on the distance from the source to the destination storage. The reward from the “random walk” scheme is the lowest, because a large number of data blocks are not delivered to the destination storage by the time limit. The reward of the proposed scheme and “data network only” grows first but decreases when the time slot length is greater than 2.5 min, indicating that there exists an optimal time slot length to maximize the reward.

The system reward performance is shown in Fig. 3. Obviously, our proposed scheme achieves the highest system reward, and the reward of our scheme and “connected vehicle network only” schemes grows with longer acceptable delay of the data blocks, due to which more data blocks can arrive at the destination within the delay constraint. However, the “random walk” scheme fails to obtain higher reward, because longer delay constraint only leads to more hops before dropping a data block and higher cost as a result.

We can draw the similar conclusion from Fig. 4 that the proposed scheme significantly improves the system reward performance compared to all other three schemes. In this figure, the average vehicle density w_{veh} that ranges from 10 to 50 per km is obtained by $L_{\text{road}}/N_{\text{veh}}$, where L_{road} is a constant and N_{veh} varies.

Key Applications

Offloading delay-tolerable data traffic to connected vehicle networks is one of the key technologies in heterogeneous networks of Internet and vehicular networks. It efficiently utilizes the potential resource from connected vehicles and mitigates the congestion problem in data networks without extra infrastructure/hardware deployment.

Cross-References

- [Data Dissemination in Delay Tolerant Networks with Geographic Information](#)

- [Mobile Data Offloading via D2D-Assisted Communications in Wireless Networks](#)
- [Routing for Vehicular Ad Hoc Networks](#)
- [Secure Routing in Cognitive Radio Vehicular Ad Hoc Networks](#)

References

- Habachi O, Shiang HP, Van der Schaar M, Hayel Y (2013) Online learning based congestion control for adaptive multimedia transmission. *IEEE Trans Signal Process* 61(6):1460–1469
- Isento J, Rodrigues J, Dias J, Paula M, Vinel A (2013) Vehicular delay-tolerant networks – a novel solution for vehicular communications. *IEEE Intell Transp Syst Mag* 5(4):10–19. <https://doi.org/10.1109/MITS.2013.2267625>
- Kenney J (2011) Dedicated short-range communications (DSRC) standards in the United States. *Proc IEEE* 99(7):1162–1182. <https://doi.org/10.1109/JPROC.2011.2132790>
- Koo I, Yang JR, Kim K (2007) Erlang capacity analysis of CDMA systems supporting voice and delay-tolerant data services under the delay constraint. *IEEE Trans Veh Technol* 56(4):2375–2385. <https://doi.org/10.1109/TVT.2007.897655>
- Krishnamurthy V (2011) Bayesian sequential detection with phase-distributed change time and nonlinear penalty – a POMDP lattice programming approach. *IEEE Trans Inf Theory* 57(10):7096–7124. <https://doi.org/10.1109/TIT.2011.2165152>
- Padmanabhan S, Stephen R, Murthy C, Coupechoux M (2015) Training-based antenna selection for PER minimization: a POMDP approach. *IEEE Trans Commun* 63(9):3247–3260. <https://doi.org/10.1109/TCOMM.2015.2455504>
- Si P, Yu FR, Ji H, Leung VC (2009) Distributed sender scheduling for multimedia transmission in wireless mobile peer-to-peer networks. *IEEE Trans Wirel Commun* 8(9):4594–4603
- Szymanski TH (2013) Max-flow min-cost routing in a future-Internet with improved QoS guarantees. *IEEE Trans Commun* 61(4):1485–1497
- Tornell S, Calafate C, Cano JC, Manzoni P (2015) DTN protocols for vehicular networks: an application oriented overview. *IEEE Commun Surv Tutor* 17(2):868–887. <https://doi.org/10.1109/COMST.2014.2375340>
- Uhlemann E (2015) Introducing connected vehicles. *IEEE Veh Technol Mag* 10(1):23–31. <https://doi.org/10.1109/MVT.2015.2390920>
- Young S, Gasic M, Thomson B, Williams J (2013) POMDP-based statistical spoken dialog systems: a review. *Proc IEEE* 101(5):1160–1179. <https://doi.org/10.1109/JPROC.2012.2225812>
- Zhang X, Cheng W, Zhang H (2014a) Heterogeneous statistical QoS provisioning over 5G mobile wireless networks. *IEEE Netw* 28(6):46–53. <https://doi.org/10.1109/MNET.2014.6963804>
- Zhang Y, Zhang Y, Qin S, Li B, He Z (2014b) Delay-bounded priority-driven resource allocation for video transmission over multihop networks. *IEEE Trans Circuits Syst Video Technol* 24(7):1184–1196. <https://doi.org/10.1109/TCSVT.2014.2302511>
- Zhao Q, Tong L, Swami A, Chen Y (2007) Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework. *IEEE J Sel Areas Commun* 25(3):589–600. <https://doi.org/10.1109/JSAC.2007.070409>

Oligopoly Pricing in Wireless Networks

Carlee Joe-Wong¹, Liang Zheng², and Jiasi Chen³

¹Electrical and Computer Engineering, Carnegie Mellon University, Silicon Valley Campus, Moffett Field, CA, USA

²Princeton University, Princeton, NJ, USA

³University of California, Riverside, CA, USA

Synonyms

[Heterogeneous network pricing](#); [Multi-provider pricing](#)

Definition

An oligopoly refers to a market in which only a few (typically two to four) companies compete with each other to offer a resource or service to users. Such market structures typically arise in industries that are difficult for companies to enter, e.g., due to government regulations or the need for significant initial investment. In wireless networks, oligopolies can be found in markets for wireless connectivity, i.e., selling users access to wireless networks. In the cellular connectivity market, for instance, cellular providers generally need government spectrum licenses to operate, and significant investment is needed to build out a cellular network. Since spectrum is limited and rarely auctioned off, new providers may

find it hard to obtain spectrum access and break into the market. In the USA today, this market is dominated by Verizon, AT&T, Sprint, and T-Mobile.

Historical Background

Wireless service providers have sold their users access to cellular data since the 1990s, when second-generation (2G) wireless networks first became available. In the mid-2000s, wireless speeds increased with the release of first 3G and then 4G networks. At the same time, users' demand for wireless data began to increase exponentially, fueled by the release of the iPhone in 2007 and the growing popularity of smartphones. This popularity in turn drove cellular providers to further upgrade their networks, with LTE introduced around 2010 in North America, Europe, and Asia. These complementary trends have contributed to large growth in the number of subscribers as well as the volume of wireless data used by existing subscribers. Yet in the midst of this growth, cellular providers in the USA and other countries attempted to dampen competition by requiring users to purchase 2-year contracts and "locked" phones that would only function on a single provider's network. In doing so, they created effective "parallel monopolies," with few users switching between providers (Joe-Wong et al. 2015). Since 2016, however, cellular providers in the USA have begun to offer unlocked phones to users, allowing them to more freely switch between providers and opening the market to more oligopolistic competition.

As demand for wireless data grew, the limit on publicly available spectrum for cellular connectivity created congestion on cellular networks, fueling an opportunity for new types of business models. Many providers showed significant interest in pricing mechanisms to incentivize changes in user demand patterns, for instance, offering users discounts at uncongested times in order to incentivize them to shift their usage from congested to uncongested times (Sen et al. 2015). Cellular providers also began to

explore offloading data usage to other types of wireless networks. WiFi, for instance, is now a near-ubiquitous presence in urban environments. Much work has been devoted to the idea of using WiFi offloading as a solution to cellular network congestion, with some studies showing that such offloading has significant potential to alleviate congestion (Lee et al. 2013). Indeed, wireless networks encompass much more than cellular networks; they include not only WiFi technology but also femtocells, Bluetooth, etc. Such technologies have become increasingly important for the Internet of things (IoT) as more devices gain wireless connectivity (Wang et al. 2017).

As some providers explored offloading and pricing solutions to growing network congestion, newer providers began to enter the market, promising users better service at lower prices. Virtual Internet service providers (ISPs), sometimes known as MVNOs (mobile virtual network operators), rent physical capacity from a single cellular provider and resell it to end users, often at a discount. In recent years, some virtual ISPs have begun combining resources on cellular networks managed by different ISPs to off-load data between them. Competition from these virtual ISPs promises to change the structure of the wireless data market.

Foundations

Traditional oligopoly models from economic theory serve as a useful starting point for investigating oligopoly pricing in wireless networks. However, wireless networks have unique characteristics that are generally not found in oligopolies. In particular, users on such networks experience externalities from other users, e.g., due to congestion. These negative externalities affect the nature of competition between wireless networks and may encourage users to utilize multiple networks in order to avoid congestion (Cheung and Huang 2015). Moreover, some users can subscribe to multiple wireless networks at once, e.g., cellular and WiFi networks. This effect is particularly

important when modeling offloading scenarios, in which competing wireless network providers may provide different services. For instance, WiFi often has higher throughput than LTE networks but generally has a smaller coverage area. In economic terms, these providers offer differentiated products.

Oligopoly models: Oligopoly pricing can often be modeled as the outcome of a game. The three most common models are called the Cournot, Bertrand, and Stackelberg models. The Cournot and Bertrand models assume that there are two existing providers offering a homogeneous product (i.e., users do not intrinsically prefer one provider over the other) and consider providers that, respectively, compete on the basis of product quantity and price. The Stackelberg model considers the behavior of an incumbent provider and a potential competitor that wishes to enter the market (Sen et al. 2013). In each model, the traditional analysis assumes that all providers attempt to maximize their individual profits. The analysis then focuses on finding the Nash equilibrium, a point at which no provider can increase its profit by changing its decisions.

The Cournot and Bertrand models are rarely used in a wireless context, as they do not capture the externality or coverage characteristics of wireless networks. Stackelberg models, however, can be used to model the entry of heterogeneous wireless technologies into the market, e.g., a provider introducing femtocells (Ren et al. 2013). In a Stackelberg model, an incumbent provider can decide which prices to charge for its services, knowing that another provider may later attempt to enter the market. This provider is called the leader of the game, since it moves first. The entrant provider can then respond to the leader; it is called the follower of the game. Since the leader can anticipate the follower's decision, it often has more market power. Stackelberg games are solved with *backward induction*, i.e., first solving for the follower's optimal actions or best response as a function of the leader's action and then solving for the leader's optimal action given the follower's best response.

Modeling product differentiation: Wireless networks can encompass many different technologies with different characteristics. Variants of a Hotelling location model are often used to capture this product differentiation. In such models, users are assumed to lie along a continuum between two providers; they experience a cost of choosing each provider that is proportional to their distance to the provider. In a wireless context, this continuum can model physical distances between a user and a base station or the cost of switching to another provider (e.g., the inconvenience of signing up for a new data plan). This virtual or physical distance between providers models users' intrinsic preferences for different providers. Users then decide to subscribe to a provider based on these intrinsic preferences, the quality of service and coverage offered, and the prices charged by each provider.

Key Applications

Wireless offloading models: Variants of the Hotelling model can be used to study the role of pricing in wireless offloading (Joe-Wong et al. 2015). In this case, users might choose between subscribing to a cellular network, a cellular network supplemented by WiFi hotspots, or no network at all. Users may be modeled as lying on a continuum between subscribing to a network or not subscribing; different users would then make different subscription choices. A key feature of such a network is that due to externalities (i.e., more users leading to more congestion), the number of users on the network influences users' satisfaction from using the network. Thus, users' decisions would be influenced by not only their location along the continuum and the prices charged but also by the number of other users subscribing to each provider. As a greater proportion of users subscribes to the cellular provider, for instance, congestion will increase, driving users to subscribe to the WiFi-supplemented network. One can then show that these dynamics

reach an equilibrium and that the prices offered can influence the equilibrium number of users on each network (Joe-Wong et al. 2015).

Stackelberg game models have also been used to analyze subscription decisions for heterogeneous wireless technologies (Duan et al. 2014). In such models, a cellular or WiFi provider might play the role of a leader, with users playing the role of followers and deciding whether to subscribe to the provider and how much data they will use. The provider can then decide which types of pricing to offer, given these user decisions. Variants of the model include having global and local WiFi providers negotiate with each other to provide a larger WiFi coverage area for users.

Other works in this space have shown significant economic benefits to offloading with a single cellular provider (Lee et al. 2014). Subscription models, however, capture only the months-long timescale of users' subscription decisions. Another aspect to oligopoly pricing in the context of heterogeneous networks is users' shorter-term decisions of when to use which type of network, given that they sometimes have access to WiFi hotspots. This switching may be complicated by the energy expenditures required to switch between networks, as well as different energy usage in cellular and WiFi networks (Yu et al. 2016). These decisions may also incorporate pricing considerations, e.g., if users have a monthly cap on their cellular data usage and must decide when to off-load data throughout the month so as not to exceed this cap (Im et al. 2016). Such models require going beyond traditional oligopoly pricing and may incorporate dynamic programming or other online optimization methods.

Virtual Internet service providers: Recent innovations in the wireless data market have led to the creation of virtual ISPs, which combine resources from multiple physical ISPs. Stackelberg models can be used to model the resulting economic interactions: the virtual ISP can act as a leader of the game, deciding how much to charge users and which providers to rent capacity from. Physical providers, knowing these

decisions, can then play the role of followers and decide whether to partner with the virtual ISP. Finally, users can follow these provider decisions when deciding whether to subscribe to the virtual ISP (Zheng et al. 2017a). Such models can also be generalized to include WiFi connectivity as one of the features of a virtual ISP data plan (Zheng et al. 2017b).

Future Directions

With the demise of fixed-term contracts and the rise of unlocked devices in the USA, we expect to see more oligopolistic wireless markets in the coming years. These measures are already fueling more competition between providers, with some even offering discounts for new users who defect from other providers. We may also soon see more providers offering data plans that combine the physical resources of multiple providers. Indeed, in countries like India, many users already use multiple SIM cards, allowing them access to multiple providers.

Oligopoly pricing in wireless networks will also be heavily influenced by the Internet of things' needs for wireless connectivity. As these devices proliferate, with unique traffic patterns that are distinct from those of user requests on traditional cellular and WiFi networks, they will likely give rise to new pricing models that will require new analyses of wireless heterogeneity and oligopolistic competition in the wireless market.

Cross-References

- [Market Entry](#)
- [Preferences and Utility Functions](#)

References

- Cheung MH, Huang J (2015) DAWN: delay-aware Wi-Fi offloading and network selection. *IEEE J Sel Areas Commun* 33(6)

- Duan L, Huang J, Shou B (2014) Pricing for local and global Wi-Fi markets. *IEEE Trans Mob Commun* 14(5):1056–1070.
- Im Y, Joe-Wong C, Ha S, Sen S, Kwon TT-K, Chiang M (2016) AMUSE: empowering users for cost-aware offloading with throughput-delay tradeoffs. *IEEE Trans Mob Commun* 15(5):1062–1076.
- Joe-Wong C, Sen S, Ha S (2015) Offering supplementary network technologies: adoption behavior and offloading benefits. *IEEE/ACM Trans Networking* 23(2):355–368.
- Lee K, Lee J, Yi Y, Rhee I, Chong S (2013) Mobile data offloading: how much can WiFi deliver? *IEEE/ACM Trans Networking* 21(2):536–550.
- Lee J, Yi Y, Chong S, Jin Y (2014) Economics of WiFi offloading: trading delay for cellular capacity. *IEEE Trans Wirel Commun* 13(3):1540–1554.
- Ren S, Park J, van der Schaar M (2013) Entry and spectrum sharing scheme selection in femtocell communications markets. *IEEE/ACM Trans Networking* 21(2):218–232.
- Sen S, Joe-Wong C, Ha S, Chiang M (2013) Smart data pricing: economic solutions to network congestion. *SIGCOMM E-book on Recent Advances in Networking*, eds. Hamed Haddadi and Olivier Bonaventure. Available at: http://sigcomm.org/education/ebook/SIGCOMMBook2013v1_chapter5.pdf
- Sen S, Joe-Wong C, Ha S, Chiang M (2015) Smart data pricing: using economics to manage network congestion. *Commun ACM* 58(12):86–93.
- Wang M, Chen J, Aryafar E, Chiang M (2017) A survey of client-controlled HetNets for 5G. *IEEE Access* 5:2842–2854.
- Yu H, Cheung MH, Huang L, Huang J (2016) Power-delay tradeoff with predictive scheduling in integrated cellular and Wi-Fi networks. *IEEE J Sel Areas Commun* 34(4):735–742.
- Zheng L, Joe-Wong C, Chen J, Brinton CG, Tan CW, Chiang M (2017a) Economic viability of a virtual ISP. In *Proceedings of the IEEE Conference on Computer Communications (IEEE INFOCOM)*, IEEE, pp. 1–9
- Zheng L, Chen J, Joe-Wong C, Tan CW, Chiang M (2017b) An economic analysis of wireless network infrastructure sharing. In the *15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, IEEE, pp. 1–8

Omnidirectional Transmission

► Omnidirectional Transmission for Massive MIMO

Omnidirectional Transmission for Massive MIMO

Xin Meng¹, Xiqi Gao², and Ming Xiao³

¹Huawei Technologies Co., Ltd., Shanghai, China

²National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

³Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Massive MIMO; Omnidirectional transmission

Definition

Omnidirectional transmission is used for cell-wide coverage for downlink common signals. It guarantees equal transmission power on each spatial direction.

Background

Massive multiple-input multiple-output (MIMO) is known to offer the advantages in increasing the spectrum and energy efficiency significantly, even with very low-complexity linear signal processing. It has received considerable interest from both academia and industry and is considered as one of the main techniques in the fifth generation (5G) of cellular systems (Marzetta 2010; Larsson et al. 2014).

In the context of massive MIMO, the transmission approaches for dedicated channels engaged in the exchange of user-specific data have been investigated intensively (Ngo et al. 2013; Adhikary et al. 2013; Sun et al. 2015; You et al. 2015; Lu et al. 2016). However, the transmission approaches for public channels have received little attention. Public channels

play crucial roles in cellular systems. Many essential common messages or services are provided to users by the base station (BS) through public channels, e.g., synchronization signals, reference signals, control signals, and multimedia broadcast/multicast services. Since public channels serve all the users in the cell rather than only certain active users, and not all the users' channel state information (CSI) can be obtained by the BS, the BS has to transmit common signals in public channels omnidirectionally to ensure cell-wide coverage. Hence the user in any spatial direction can have a reliable receiving signal-to-noise ratio (SNR).

Most omnidirectional transmission schemes developed for the conventional small-scale MIMO systems, such as single-antenna transmission or cyclic delay diversity (CDD), may be incompatible with massive MIMO. For example, consider single-antenna transmission, where one single antenna is selected from BS antennas to broadcast signals. In this case, the selected antenna must be equipped with a much more powerful and expensive power amplifier (PA) to ensure the same coverage area as all antennas are used for broadcasting signals. Nonetheless, this scheme would not work in a massive MIMO system. Since by using a large number of antennas, each antenna is made small with a much less powerful PA, and no antenna would be sufficiently powerful to serve as the single antenna mentioned above. Therefore, it is necessary to design an omnidirectional transmission approach for massive MIMO.

Omnidirectional Precoding-Based Transmission

It is known that without assuming any spatial structure for the channel, the length of the downlink pilot should not be less than the number of BS antennas. When considering massive MIMO where the number of BS antennas is very large, such a huge pilot overhead will lower the net spectrum efficiency significantly. In order to reduce the downlink pilot overhead, an omnidirectional precoding

(OP)-based transmission is proposed in Meng et al. (2016), where the high-dimension signal vector transmitted from the BS antenna array is composed by an $M \times N$ omnidirectional precoding (OP) matrix $\mathbf{W}$ and a low-dimensional signal vector, where M is the number of BS antennas. Owing to the OP matrix, the user does not need to estimate the actual channel of dimension M , but just needs to estimate the equivalent channel, i.e., the product of the actual channel and the OP matrix, of dimension N . In this case, the pilot length can be reduced to N .

Under the framework of OP-based transmission, the OP matrix should satisfy three basic requirements. Firstly, it should guarantee omnidirectional transmission, i.e., the transmission power on each spatial direction should be equal. Secondly, it should guarantee equal transmission power on each antenna to sufficiently utilize all the available PA capacities of BS antennas. Finally, it should maximize the ergodic rate for the Rayleigh independent and identically distributed (i.i.d.) channel. These three requirements can be, respectively, equivalent to three mathematical conditions:

$$\text{diag}(\mathbf{F}_M \mathbf{W} \mathbf{W}^H \mathbf{F}_M^H) = \frac{N}{M} \mathbf{1}_M \quad (1)$$

$$\text{diag}(\mathbf{W} \mathbf{W}^H) = \frac{N}{M} \mathbf{1}_M \quad (2)$$

$$\mathbf{W}^H \mathbf{W} = \mathbf{I}_N, \quad (3)$$

where $\mathbf{F}_M$, $\mathbf{1}_M$, and $\mathbf{I}_M$ denote the unitary M -point discrete Fourier transform (DFT) matrix, the $M \times 1$ column vector of all ones, and the $M \times M$ identity matrix, respectively, and $\text{diag}(\mathbf{A})$ denotes a column vector constituted by the main diagonal of a square matrix $\mathbf{A}$. Conditions (1), (2), and (3) mean that all the row vectors of $\mathbf{F}_M \mathbf{W}$ should have the same 2-norm, all the row vectors of $\mathbf{W}$ should have the same 2-norm, and $\mathbf{W}$ itself should be semi-unitary, respectively.

The OP matrix satisfying the above three conditions simultaneously can be designed with two different methods. The first one is based on Zadoff-Chu (ZC) sequences. A ZC sequence of length M has an elegant property that all

the M elements in this sequence have the same amplitude and all the M elements of the M -point DFT of this sequence still have the same amplitude (Chu 1972). By utilizing this property, the N columns of an $M \times N$ OP matrix $\mathbf{W}$ can be set as different cyclically shifted versions of a ZC sequence or different linearly modulated versions of a ZC sequence. In addition, when M is an integer multiple of N , $\mathbf{W}$ can also be constructed as

$$\mathbf{W} = \sqrt{\frac{N}{M}} \text{diag}(\mathbf{c})(\mathbf{1}_{M/N} \otimes \mathbf{I}_N) \quad (4)$$

where $\mathbf{c}$ denotes a ZC sequence of length M , $\text{diag}(\mathbf{c})$ denotes the diagonal matrix with $\mathbf{c}$ on the main diagonal, and $\otimes$ denotes the Kronecker product. The OP matrix in (4) not only satisfies conditions (1), (2), and (3) simultaneously but also has the following advantages: (1) it asymptotically maximizes the ergodic rate, maximizes the diversity order, and minimizes the outage probability in the large-scale array regime, not only for the i.i.d. channel but also for spatially correlated channels; (2) it preserves the peak-to-average power ratio (PAPR) of the transmitted signal after precoding (Meng et al. 2016). The second method for OP matrix design is based on Golay complementary sequences. A Golay complementary pair consists of two binary sequences with the same length and satisfying the complementary property that the respective aperiodic autocorrelation functions of these two sequences sum to a Kronecker delta function (Golay 1961). When utilizing Golay complementary sequences for OP matrix design, the number N of the columns of an $M \times N$ OP matrix $\mathbf{W}$ should be equal to or greater than two. This is different from the case with ZC sequence-based design, where N can be equal to one. When $N = 2$, the two columns of an $M \times N$ OP matrix $\mathbf{W}$ can be set as the two sequences in a Golay complementary pair directly. When $N > 2$ and N is even, the N columns of an $M \times N$ OP matrix $\mathbf{W}$ can be set as N -specific columns of an $M \times M$ Golay-Hadamard matrix (Meng et al. 2018). One advantage of Golay complementary sequence-based OP matrix design is that it guarantees equal

transmission power on any continuous spatial direction, while the ZC sequence-based design only guarantees equal transmission power on discrete spatial directions. In addition, since Golay complementary sequences are binary sequences, i.e., the elements take values in $\{\pm 1\}$, another advantage of Golay complementary sequence-based OP matrix design is that it will be suitable for radio-frequency (RF) precoding in the analog domain in millimeter wave systems (Xiao et al. 2017; Huang et al. 2018, to appear).

Conclusions

Omnidirectional transmission is used to guarantee cell-wide coverage for downlink common signals. To reduce the downlink pilot overhead in massive MIMO systems, the high-dimensional signal vector transmitted from the BS antenna array is composed by an OP matrix and a low-dimensional signal vector. The OP matrix has to satisfy three basic requirements, including omnidirectional coverage, equal transmission power on each antenna, and ergodic rate maximization for the Rayleigh i.i.d. channel. These three requirements are equivalent to three mathematical conditions correspondingly. There are two different methods to design such an OP matrix satisfying the three conditions simultaneously, based on ZC sequences and Golay complementary sequences, respectively.

Key Applications

Precoding Methods for Downlink Public Channels

References

- Adhikary A, Nam J, Ahn J, Caire G (2013) Joint spatial division and multiplexing—the large-scale array regime. *IEEE Trans Inf Theory* 59(10):6441–6463

- Chu D (1972) Polyphase codes with good periodic correlation properties. *IEEE Trans Inf Theory* 18(4): 531–532
- Golay MJE (1961) Complementary series. *IRE Trans Inf Theory* 7(2):82–87
- Huang Y, Zhang J, Xiao M (2018, to appear) Constant envelope hybrid precoding for directional millimeter-wave communications. *IEEE J Sel Areas Commun*
- Larsson EG, Edfors O, Tufvesson F, Marzetta TL (2014) Massive MIMO for next generation wireless systems. *IEEE Commun Mag* 52(2):186–195
- Lu AA, Gao XQ, Zheng YR, Xiao C (2016) Low complexity polynomial expansion detector with deterministic equivalents of the moments of channel gram matrix for massive MIMO uplink. *IEEE Trans Commun* 64(2):586–600
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wirel Commun* 9(11):3590–3600
- Meng X, Gao XQ, Xia XG (2016) Omnidirectional precoding based transmission in massive MIMO systems. *IEEE Trans Commun* 64(1):174–186
- Meng X, Gao XQ, Xia XG (2018) Omnidirectional precoding and combining based synchronization for millimeter wave massive MIMO systems. *IEEE Trans Commun* 66(3):1013–1026
- Ngo HQ, Larsson EG, Marzetta TL (2013) Energy and spectral efficiency of very large multi-user MIMO systems. *IEEE Trans Commun* 61(4):1436–1449
- Sun C, Gao XQ, Jin S, Matthaiou M, Ding Z, Xiao C (2015) Beam division multiple access transmission for massive MIMO communications. *IEEE Trans Commun* 63(6):2170–2184
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis GK, Björnson E, Yang K, I CL, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun* 35(9):1909–1935
- You L, Gao XQ, Xia XG, Ma N, Peng Y (2015) Pilot reuse for massive MIMO transmission over spatially correlated Rayleigh fading channels. *IEEE Trans Wirel Commun* 14(6):3352–3366

On-Demand WLAN

- [Radio-on-Demand Wireless LAN](#)

One-Way Relay

- [Fundamental Properties of Wireless Relays and Their Channel Estimation](#)

Operation and Maintenance of Data Centers

- [Data Center Networking \(DCN\): Infrastructure and Operation](#)

Opportunistic Data Forwarding

- [Opportunistic Routing \(OR\)](#)

Opportunistic Encounter

- [Delay-Tolerant Network Routing](#)

Opportunistic Networks

- [Data Dissemination in Delay Tolerant Networks with Geographic Information](#)

Opportunistic Routing

- [Joint Network Coding and Opportunistic Forwarding](#)

Opportunistic Routing (OR)

Yuanzhu Chen

Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Synonyms

[Any-path routing](#); [Cooperative diversity routing](#); [Diversity forwarding](#); [Opportunistic data forwarding](#)

Definition

The basic idea of opportunistic routing is to utilize all downstream nodes as potential forwarders and coordinate transmissions among forwarders. It has ability to overhear the transmitted packet and to coordinate among relaying nodes. In OR, the next-hop forwarder is decided after the sender broadcasts the data packets. Multiple downstream nodes in addition to that matching receiver will receive these packets and be potential forwarders. One of these potential forwarders, which is “closest” to the destination, will be selected to forward packets. The on-the-fly selection of the next-hop is the fundamental principle of opportunistic routing. Since multiple downstream nodes are potential next-hop forwarders, opportunistic routing can reduce the possibilities of reconstructing paths or retransmitting packets due to link breakage on a preselected path. Therefore, opportunistic routing benefits from the broadcast characteristic of wireless mediums to improve network performance.

Historical Background

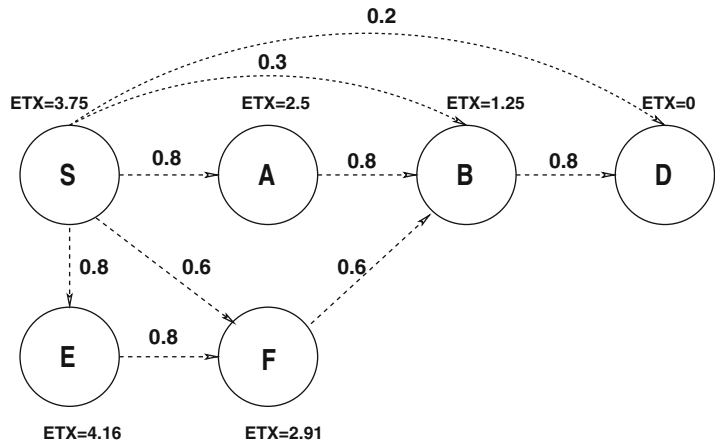
A multi-hop wireless mesh network, wireless mesh in short, is a wireless data communication network in which end-to-end nodes are situated beyond direct radio transmission ranges. Intermediate nodes are required to forward data in order to enable communication between nodes that are far apart. Compared to single-hop wireless networks, such as Wi-Fi, Bluetooth, and cellular networks, multi-hop wireless networks can have extended communication ranges. Furthermore, they can provide coverage in hard-to-wire areas and improve the network throughput by short distance with reliable transmissions. The distinct features and critical design factors of multi-hop wireless networks also pose many problems. Routing is a central problem in wireless mesh. Routing algorithms will select a subset of intermediate nodes to create a path such that these intermediate nodes can forward packets from the source to the destination.

Most routing solutions are called traditional routing because they follow the Internet architecture of determining routes before forwarding data packets along the routes. Traditional routing preselects a fixed routing path before the source node starts transmissions. Each node in a route uses a preselected neighbor as the next-hop to forward data packets towards the destination. Once all next-hops have been selected, all packets transmitted between the source and the destination follow the same path. These routing solutions borrow ideas from the routing protocols for wire-line networks and do not adapt well to the dynamic wireless environment where transmission failure occur frequently. Wireless mesh networks are built on top of wireless links, and no longer have the constraints of point-to-point links prevalent in the Internet. In particular, wireless links are broadcast and subject to channel fading, which causes an uncertain number of nodes to receive a single transmission. In wireless networks, when a node transmits a data packet, multiple nodes can overhear the transmission, which can compensate for the transmissions failure to the preselected next-hop. However, traditional routing cannot utilize this broadcast nature of wireless mesh, because it dictates that all nodes without a matching receiver address should drop packets, and only the node that the routing module decides to be the next hop can keep packets for subsequent forwarding (Chakchouk 2015).

Foundations

The idea behind opportunistic routing is to utilize multiple nodes as potential forwarders and exploit multiple long but may unstable links concurrently, resulting in high expected progress per transmission (Biswas and Morris 2005). The routing for packet transmissions is not predetermined and can be updated for each packet of a flow. In opportunistic routing, a set of intermediate nodes between a source and a destination will be selected as potential forwarders, which are also called candidate nodes. Candidate selection algorithms decide the candidate nodes and determine the proper priority for each candidate.

Opportunistic Routing (OR), Fig. 1 An example of opportunistic routing



An effective routing metric will be used for the selection of candidate nodes and ordering nodes by their priorities (Zhong et al. 2006). The routing metric can show the ability of nodes to act as the next forwarder. The highest ability of a node may indicate that this node is “closest” to the destination. Initially, the typical routing metrics are from traditional routing protocols, for example, the expected number of transmissions or the hop count metric. After the selection of candidate nodes, opportunistic routing allows any potential forwarder to involve in packet transmissions according to the coordination algorithm, which is used to avoid duplicate transmissions and also guarantee reliable transmissions.

With a coordination algorithm, nodes that have received data packets from upstream nodes will decide whether to forward packets, discard packets, or wait for others’ transmissions. To coordinate transmissions, a node can send extra coordination messages or combine coordination information into data packets. The coordination message informs other nodes whether it has received packets or not; it is an accurate way to make coordination between nodes given that the coordination message is not lost. However, extra coordination messages require network resources and will slow down the transmission of data packets because data packets are usually sent out after a node receiving the coordination message. In addition to sending extra coordination messages, most opportunistic routing protocols combine coordination messages with data

packets instead. To increase the successful receiving possibility of coordination information, opportunistic routing protocols transmit a batch of data packets together in a round with a same coordination message. The coordination messages can be received if one of data packets in that batch is received (Boukerche and Darehshoorzadeh 2015).

We use the example in Fig. 1, inspired by Extremely Opportunistic Routing (ExOR) (Biswas and Morris 2005), to elaborate how opportunistic routing works. It presents the process of opportunistic routing and shows its advantages compared to traditional routing. In this example, the packet delivery probability is shown over the link. Traditional routing chooses the shortest path, $S-A-B-D$, to transmit packets. When S sends packets to node A , data packets are transmitted by broadcasting. Downstream nodes (B or D) may receive some packets. Neither node B nor D is selected as the next hop of the source; even though they receive some packets, these packets have to be retransmitted by node A . Therefore, the traditional routing scheme has many unnecessary transmissions and wastes network resources. By contrast, opportunistic routing can increase the performance of packet transmissions by taking advantage of the broadcast nature of wireless media, specifically long-distance transmissions. ExOR uses the Expected Transmission Count (ETX (Couto et al. 2003)) as the routing metric, which calculates the average number of transmission that a packet

required to be received by the destination. The ETX value of a corresponding link is the inverse of the packet delivery probability of that link. The ETX value on a path can be derived by adding up all ETX values of each link on this path. The definition of an ETX value of a node is the minimum ETX value of paths from this node to the destination. The intermediate nodes with an ETX value smaller than that of the source will be selected as candidate nodes. In this example, the candidate nodes are $\{A, B, F, D\}$. Each intermediate node will be assigned a priority based on its ETX value. Nodes with a small ETX value will be assigned a high priority. Therefore, the order of the priority is $D > B > A > F$. After the source node sends packets, multiple candidate nodes are involved in receiving and forwarding packets. In ExOR, the high priority node will forward the received packet first, e.g., if both nodes A and B receive a packet from the source, node B will forward this packet first. When node A overhears node B 's transmission, it discards the same packet since node B already forwards this packet to downstream nodes. If node D receives this packet from the source, neither node A or B will forward the same packet.

Therefore, opportunistic routing is likely to increase total network capacity as well as individual connection throughput. It transmits each packet fewer times than traditional routing, which should cause less interference for other users of the network and of the same spectrum.

Key Applications

Opportunistic routing is fundamental in all wireless mesh networks or ad hoc networks. It has increased the efficiency, throughput, and reliability by utilizing the broadcasting nature of the wireless medium.

Cross-References

- [Ad Hoc Networks](#)
- [Network Coding](#)
- [Wireless Sensor Networks](#)

References

- Biswas S, Morris R (2005) ExOR: opportunistic multi-hop routing for wireless networks. *ACM SIGCOMM Comput Commun Rev* 35(4):133–144
- Boukerche A, Darehshoorzadeh A (2015) Opportunistic routing in wireless networks: models, algorithms, and classifications. *ACM Comput Surv* 47(2):22–36
- Chakchouk N (2015) A survey on opportunistic routing in wireless communication networks. *IEEE Commun Surv Tutorials* 17(4):2214–2241
- Couto DD, Aguayo D, Bicket J, Morris R (2003) A high-throughput path metric for multi-hop wireless routing. In: *Proceedings of the ACM international conference on mobile computing and networking*, New York, pp 134–146
- Zhong Z, Wang J, Nelakuditi S (2006) On selection of candidates for opportunistic any path forwarding. *ACM SIGMOBILE Mob Comput Commun Rev* 10(4):1–2

Opportunistic Routing in Underwater Sensor Networks

Liang Zhao¹ and Ammar Hawbani²

¹School of Computer Science, Shenyang Aerospace University, Shenyang, China

²University of Science and Technology of China, Hefei, China

Synonyms

[Acoustic networks](#); [Acoustic sensor networks](#); [Acoustic wireless sensor networks](#)

Definition

Underwater wireless sensor networks (UWSNs) are normally used to collect data from the acoustic environment and transfer the data from the sensor devices to the sonobuoys on the surface. In UWSNs, opportunistic routing (OR) (Larsson 2001) is an ideal routing paradigm to reduce energy consumption and increase path reliability.

Historical Background

Knowing and investigating the ocean is important to people in order to sustain human life. The wireless communication-enabled sensor devices can be deployed to monitor and observe aquatic information. Due to the ocean environment, the sensor nodes are mobile as freely floating by the ocean current where the nodes are connected ad hoc to form a dynamic wireless network.

Similar to other types of wireless sensor networks, most sensor nodes in UWSNs are also energy-constrained; thereby we can refer UWSNs as a special type of WSN. However, compared to traditional WSNs, UWSNs could suffer low path reliability and energy problem because of the rapid attenuation of radio signals and the dynamic movement of sensor nodes.

Also, because traditional wireless routing protocols require regular updates and maintenance in order to establish the routing paths where such routing paradigms could demand high energy consumption, this leads that the routing design of traditional routing protocols could not work efficiently for UWSNs. We can see from this point, although UWSNs and traditional wireless sensor networks share many similarities, there are still quite many distinguishing characteristics of UWSNs such as float with current, energy constrained, and narrow bandwidth.

Compared to traditional routing where packets are transmitted along predetermined paths, OR uses a prioritization metric (routing metric) to select a set of candidates as potential forwarders (usually called candidate set, CS) in OR. According to some routing criteria or priority (determines the ability to reach the destination with lower forwarding cost), the group of nodes then coordinate among each other and decide one node with the highest priority from this group to be next forwarder. This forwarder will further broadcast the data message (packet) to its neighbor to repeat the above procedure until the destination receiver receives the message. As described above, OR shows its distinguishing characteristics to work in unreliable environments. It provides efficient routing paths with the increased possibility of accomplishing the transmissions.

More, by creating strong virtual links and minimizing the number of retransmissions, the path reliability and the energy consumption are both reduced. Therefore, OR is a perfect paradigm to overcome the limits of bandwidths, channels, and energy in UWSNs.

In UWSNs, energy efficiency is the primary challenging issue mostly due to the attenuation of frequent changing topology and acoustic radio signal in the underwater environment. Hence, large variable latency and serious path loss (could be the temporary loss of paths), as well as low bandwidth, could be experienced under such condition. Further, due to the unreliable path, packet retransmission can occur with higher energy consumption. Therefore, as the energy is one of the most important issues in UWSNs, the design of the routing scheme should consider how to enable energy-efficient routing selection or packet forwarding. Besides, since sensor nodes float freely with the ocean current movement, the topology of UWSN could be highly dynamic. Forwarding the packet to a relay node with geographical advantages is essential to reducing latency. Therefore, the design of geographic-based routing schemes is to select a forwarder node in a better position (e.g., closer to the destination) in order to shorten the delay during the packet transmission.

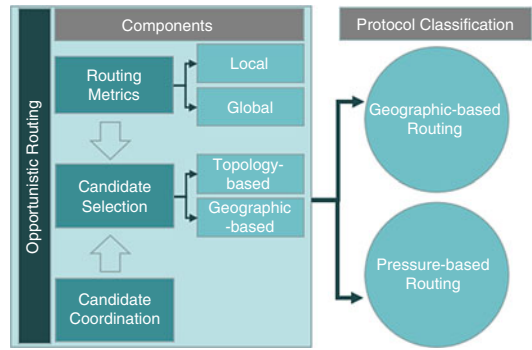
Therefore, in the following paragraphs, we introduce how OR works in detail. As shown in Fig. 1, an OR paradigm consists of three main components, routing metrics, candidate selection, and candidate coordination (Darehshoorzadeh and Boukerche 2015; Coutinho et al. 2016). Each component is presented as follows.

- (a) **Routing Metrics:** Generally, in networking, routing metrics is referred to as the method to determine whether a node or a link is better than another or not. In UWSNs, the metric could be either about the local information or about the global information. For global (or end-to-end) metric (Darehshoorzadeh and Boukerche 2015), the information of the whole network should be collected to judge the quality of a next forwarder in CS. In contrast, for local metric, only local information including the information of

CS and neighbors of CS is considered to select the next-hop forwarder. It is obvious that the global metric could help to select a better forwarder than local metric. However, while using a global metric, great overhead with high energy consumption could be created while gathering the information of the entire network. Besides, in UWSNs, it is not feasible to obtain information of the entire network which is very costly.

- (b) **Candidate Selection:** In OR, a packet is sent to the neighbors (CS) of the current sender/forwarder. Based on the information of the network (i.e., routing metrics), the nodes in the CS need to select the optimal candidate as the next forwarder(s) to forward the packet toward the destination node. One or multiple nodes from the CS could be chosen as a forwarder(s), whereas the overhead may be high if multiple forwarders are used. The candidate selection mechanisms can be divided into two main categories, topology-based and geographic-based. The global routing metrics are normally applied in the topology-based candidate selection to provide the information of network topology, while the local routing metrics are used in the geographic-based candidate selection to find the forwarder in a better position for relaying packet. As topology-based candidate selection uses information of whole network, the selection of the candidates is more complex than using the local information.
- (c) **Candidate Coordinate:** Once the packets are received by a CS, the nodes of the CS should coordinate and collaborate with each other in order to decide which node out of this CS must continue to forward the packet and which nodes should discard the packet after this. The procedure of coordinating among candidate nodes is referred as candidate coordination.

As depicted in Fig. 1, in UWSNs, the OR protocols can be classified into two main categories, geographic-based and pressure-based (Darehshoorzadeh and Boukerche 2015) where



Opportunistic Routing in Underwater Sensor Networks, Fig. 1 The procedures and classification of OR

position information and pressure (depth of sensor node) information are used for these two categories for candidate selection, respectively. The representative geographic-based routing schemes are vector-based forwarding (VBF) (Xie et al. 2006), hop-by-hop VBF, and geographic and opportunistic routing with depth adjustment-based topology control for communication recovery (GEDAR) (Coutinho et al. 2014), while the representative pressure-based routing schemes are hydraulic pressure-based anycast routing (HydroCAST) (Lee et al. 2010), void-aware pressure routing (VAPR) (Noh et al. 2013), and E-PULRP (Gopi et al. 2010).

Key Applications

UWSN is a promising technology which can be applied in many underwater (ocean, rivers, and lakes) scenarios, e.g., ocean data collection, wild benthos studies, marine climate observation, off-shore exploration, military tactical surveillance, and offshore exploration. In particular, opportunistic routing could provide a pathway with less energy consumption and enhanced reliable communication to support these scenarios. The key applications of UWSNs (Felemban et al. 2015) are presented below.

- (a) **Environment Monitoring:** The efficient routing of UWSN could be applied to monitor the

water quality, habitat as well as the offshore underwater exploration, etc. For example, the application based on UWSN has been developed to monitor the water quality of rivers in India and China (Menon et al. 2012; He and Zhang 2012). The UWSN-based European-funded project, ACMENet, is constructed to collect environment of the living environment of marine life (Acar and Adams 2006). Other work (Lloret et al. 2011; Lpez et al. 2009) also employs UWSNs to the applications of UWSNs provide a cheap and real-time monitor of various underwater species. Besides, UWSNs have been deployed to monitor and investigate the underwater natural resources and infrastructures such as marine optical cables, marine oil field, marine gas pipelines, etc. as an alternate of wired sensors (Heidemmann et al. 2006; Srinivas et al. 2012; Jawhar et al. 2007; Saeed et al. 2014).

- (b) Disaster Monitoring: UWSNs have been applied to efficiently collect the disaster data and prevent disasters in the aftermath disaster scenario. There is a multitude of UWSN applications including underwater earthquake, tsunamis (Kumar et al. 2012), flood (Pasi and Bhawe 2015), oil spillage (Khan and Jenkins 2008), volcanic eruptions, etc. For example, UWSN is deployed to detect and analyze ocean pollution, oil spillage, and gas pipeline (Khan and Jenkins 2008). Also, with the application of UWSNs, the flood indicators such as water level and temperatures could be monitored in real time (Marin-Perez et al. 2012).
- (c) Military: As an economical solution, UWSNs can be deployed to clear underwater mines and detect naval submarines as well as for near- or offshore monitoring and surveillance, using autonomous underwater vehicle (AUV). For example, sensors have been deployed on the AUV or located on the side of AUV to detect and analyze the underwater mines (Rao et al. 2009). With the advanced sensing ability, in the military, another instance of application of UWSNs is to localize the target submarines (Hamilton et al. 2010).

Cross-References

- [Routing](#)
- [Network Stack for Wireless Sensor Networks](#)

References

- Acar G, Adams AE (2006) ACMENet: an underwater acoustic sensor network protocol for real-time environmental monitoring in coastal areas. Paper presented at IEEE proceedings: radar, sonar and navigation 153(4): 2006
- Coutinho RWL et al (2014) GEDAR: geographic and opportunistic routing protocol with depth adjustment for mobile underwater sensor networks. Paper presented at IEEE ICC 2014 – ad-hoc and sensor networking symposium, June 2014
- Coutinho RWL, Boukerche A, Vieira LFM et al (2016) Design guidelines for opportunistic routing in underwater networks. *IEEE Commun Mag* 54(2): 40–48
- Darehshoorzadeh A, Boukerche A (2015) Underwater sensor networks: a new challenge for opportunistic routing protocols. *IEEE Commun Mag* 53(11):98–107
- Felemban E, Shaikh FK, Qureshi UM et al (2015) Underwater sensor network applications: a comprehensive survey. *Int J Distrib Sens Netw* 2015(11):1–14
- Gopi S et al (2010) E-PULRP: energy optimized path unaware layered routing protocol for underwater sensor networks. *IEEE Trans Wireless Commun* 9(11): 3391–3401
- Hamilton MJ, Kemna S, Hughes D (2010) Antisubmarine warfare applications for autonomous underwater vehicles: the GLINT09 sea trial results. *J Field Robot* 27(6):890–902
- He D, Zhang L-X (2012) The water quality monitoring system based on WSN. Paper presented at the 2nd international conference on consumer electronics, communications and networks, Yichang, Apr 2012
- Heidemmann J, Ye W, Wills J et al (2006) Research challenges and applications for underwater sensor networking. Paper presented at the IEEE wireless communications and networking conference, Apr 2006
- Jawhar I, Mohamed N, Shuaib K (2007) A framework for pipeline infrastructure monitoring using wireless sensor networks. Paper presented at the wireless telecommunications symposium, Apr 2007
- Khan A, Jenkins L (2008) Undersea wireless sensor network for ocean pollution prevention. Paper presented at the 3rd international conference on communication systems software and middleware and workshops, Jan 2008
- Kumar P, Priyadarshini P et al (2012) Underwater acoustic sensor network for early warning generation. Paper presented at the Oceans, IEEE, Hampton Roads
- Larsson P (2001) Selection diversity forwarding in a multihop packet radio network with fading channel

- and capture. Paper presented at SIGMOBILE mobile computing communication review, 2001
- Lee U et al (2010) Pressure routing for underwater sensor networks. Paper presented at IEEE INFOCOM, Mar 2010
- Lloret J, Sendra S, Garcia M et al (2011) Group-based underwater wireless sensor network for marine fish farms. Paper presented at the IEEE GLOBECOM workshops, Houston, Dec 2011
- Lpez M, Martnez S, Gmez J et al (2009) Wireless monitoring of the pH, NH₄⁺ and temperature in a fish farm. *Procedia Chem* 1:445–448
- Marin-Perez R, Garcia-Pintado J, Omez ASG (2012) A real time measurement system for long-life flood monitoring and warning applications. *Sensors* 12(4): 4213–4236
- Menon K, Divya P, Ramesh M (2012) Wireless sensor network for river water quality monitoring in India. Paper presented at the 3rd international conference on computing communication networking technologies, July 2012
- Noh Y et al (2013) Vapr: void-aware pressure routing for underwater sensor networks. *IEEE Trans Mobile Comput* 12(5):895–908
- Pasi AA, Bhawe U (2015) Flood detection system using wireless sensor network. *Int J Adv Res Comput Sci Softw Eng* 5(2):386–389
- Rao C, Mukherjee K, Gupta S et al (2009) Underwater mine detection using symbolic pattern analysis of side scan sonar images. Paper presented at the American control conference, June 2009
- Saeed H, Ali S, Rashid S et al (2014) Reliable monitoring of oil and gas pipelines using wireless sensor network (wsn): remong. Paper presented at the 9th international conference on system of systems engineering, June 2014
- Srinivas S, Ranjitha P, Ramya R, Narendra G (2012) Investigation of oceanic environment using large-scale UWSN and UANETS. Paper presented at the 8th international conference on wireless communications, networking and mobile computing, Sept 2012

- Xie P, Cui J-H, Lao L (2006) VBF: vector-based forwarding protocol for underwater sensor networks. *Networking* 3976:1216–1221

Optical Camera Communication

- [Indoor Visible Light Communication Networks for Camera-Based Mobile Sensing](#)

Orthogonal FDM

- [Principle of OFDM and Multi-carrier Modulations](#)

Outage Probability

- [Transmission Capacity](#)

Outsider, External

- [Wireless Sensor Network Security](#)

P

Packet Forwarding Function

- [Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization](#)

PAPR Reduction in OFDM Systems

Hao Li
Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada

Synonyms

[Crest factor reduction](#); [Peak factor reduction](#)

Definitions

Peak-to-average power ratio (PAPR) reduction is a technique used to reduce the peak-to-average power ratio of transmitted signals so that the power amplifier at the transmitter is not driven into saturation by high signal peaks.

Historical Background

The development of PAPR reduction techniques dates back to 1970s, when complementary sequences were utilized to minimize the peak factor of multitone signals (Schroeder 1970; Greenstein and Fitzgerald 1981; Boyd 1986). This was the initial coding strategy to reduce the PAPR of multi-carrier signals. Along with the popularity and commercial use of orthogonal frequency division multiplexing (OFDM) in 1990s, many efforts have been made to overcome the high-PAPR challenge in OFDM transmission. Various methods of PAPR reduction in OFDM systems, as well as their subsequent improvement schemes, were proposed. In 1994, a block coding scheme was proposed (Jones et al. 1994), by adding a Simple Odd Parity Code (SOBC) at the end of each three data bits to reduce the PAPR. The simplest and most widely used clipping technique was introduced by O'Neill and Lopes in 1995 (O'Neill and Lopes 1995), which basically clipped the parts of signals that were outside a predetermined region. In 1996, the coding strategy was further developed to have both PAPR reduction and error correction capabilities (van Nee 1996). Within the same year, a selected mapping (SLM) method (Bäumel et al. 1996) and interleaving method (Eetvelt et al. 1996) were published, by exploiting the multi-carrier attribute of OFDM signals. A few months later, a partial transmit sequence (PTS) method that optimally rotated signal subblocks was proposed (Müller and Huber

1997). In 1998, a peak windowing technique that could suppress out-of-band radiation while reducing the peak signal was introduced (van Nee and de Wild 1998). Another three typical PAPR reduction methods for OFDM, including nonlinear companding transform (Wang et al. 1999), tone reservation (TR), and tone injection (TI) (Tellado 1999), were proposed in 1999.

Modifications of these key PAPR reduction strategies are continuously investigated till today, in order to achieve a trade-off between the capacity of PAPR reduction and the transmission power, data rate loss, implementation complexity, and bit-error-rate (BER) performance. In recent years, some hybrid methods were also proposed by hybridizing two or more PAPR reduction schemes (Sandoval et al. 2017).

PAPR of OFDM Signals

OFDM has been widely adopted in modern communication systems. For an OFDM signal with N subcarriers, it can be mathematically expressed as

$$x(t) = \sum_{k=0}^{N-1} X(k) e^{j2\pi f_k t}, \quad 0 \leq t \leq T, \quad (1)$$

where $X(k)$ is the data symbol transmitted on the k th subcarrier and T is the duration of one OFDM signal. $f_k = k\Delta f$ is the frequency of the k th subcarrier, where Δf is the subcarrier spacing and $T\Delta f = 1$.

Normally, data symbols to be transmitted on the subcarriers, i.e., $X(k)$, $k = 0, 1, \dots, N-1$, are independent, identically distributed (i.i.d.) random variables. Based on the central limit theorem, when N is considerably large, the resulting signal $x(t)$ approaches a complex Gaussian process. A small percentage of signal points will take very large magnitudes, resulting into a high PAPR. Technically, the PAPR of a time-domain OFDM signal $x(t)$ is defined as the ratio between its maximum instantaneous power and the average power, which can be written as

$$\text{PAPR}[x(t)] = \frac{\max_{0 \leq t \leq T} |x(t)|^2}{\frac{1}{T} \int_0^T |x(t)|^2 dt}. \quad (2)$$

In the case that a time-domain OFDM signal $x(t)$ is obtained by L -times oversampling, the PAPR computed from the L -times oversampled signal samples can be defined as

$$\text{PAPR}[x(n)] = \frac{\max_{0 \leq n \leq NL-1} |x(n)|^2}{\frac{1}{NL} \sum_{n=0}^{NL-1} |x(n)|^2}. \quad (3)$$

A large PAPR would drive power amplifiers at the transmitter into saturation, causing in-band signal distortion and out-of-band radiation. The nonlinear distortions would increase the BER and induce a serious degradation in system performance.

Key PAPR Reduction Methods in OFDM Systems

Various approaches have been proposed to reduce the PAPR in OFDM systems. In general, these methods can be classified into three categories: signal distortion techniques, signal scrambling techniques, and hybrid techniques.

Signal Distortion Techniques

Clipping and Filtering Clipping is a method to reduce the signal peaks to a desired level. If the signal exceeds a predetermined clipping level A , a clipper would limit the signal envelope to the clipping level. Otherwise, the clipper passes the signal without change (O'Neill and Lopes 1995). The output signal of the clipping operation, $\hat{x}(n)$, can be given by

$$\hat{x}(n) = \begin{cases} x(n), & |x(n)| \leq A \\ Ae^{j\angle[x(n)]}, & |x(n)| > A \end{cases}, \quad (4)$$

where $\angle[x(n)]$ represents the phase of $x(n)$. Clipping is a nonlinear process that leads to both in-band and out-of-band distortions. Filtering can

thus be employed after clipping to reduce out-of-band radiation. However, filtering may also cause peak regrowth so that the signal after clipping and filtering could exceed the clipping level again at some points. To reduce overall peak regrowth, an iterative clipping-and-filtering operation can be used at the cost of additional computational complexity (Li and Cimini 1998).

Peak Windowing Peak windowing method attenuates signal peaks in a much smoother way compared to the hard clipping, resulting in reduced distortion. In this method, the original OFDM signal is multiplied by a shaped window such as Gaussian, Kaiser, Hamming, or cosine filters (van Nee and de Wild 1998). As these windows have a narrow impulse response in the time domain and approximately rectangular spectrum in the frequency domain, peak windowing can suppress out-of-band radiation while reducing the signal peaks.

Companding Transform Companding transform is a method to nonlinearly scale the OFDM signal by suppressing signal points with large amplitudes and expanding signal points with small amplitudes (Wang et al. 1999). At the receiver, the original signal is recovered from the companded signal via the decompanding process. Different companding algorithms can be utilized, including linear symmetrical transform (LST), linear asymmetrical transform (LAST), nonlinear symmetrical transform (NLST), and nonlinear asymmetrical transform (NLAST) (Rahmatallah and Mohan 2013). For instance, the LST companding algorithm can be written as

$$\hat{x}(n) = ax(n) + b, \quad (5)$$

where $0 < a < 1$ is the slop parameter and $b > 0$ is the bias parameter. Companding transform is able to reduce PAPR in OFDM systems without any side information. However, it could increase the quantization error, resulting into a higher BER.

Signal Scrambling Techniques

Coding The principle of coding method is to reduce the occurrence probability of the same phase at OFDM subcarriers, so as to reduce the occurrence of very high peaks. Block coding, convolutional coding, and concatenate coding schemes can all be employed to reduce the PAPR (Jones et al. 1994; van Nee 1996). The main disadvantage of the coding method is the high computational complexity to search for adequate code words, particularly when the number of subcarriers is large. Also, the PAPR reduction is at the cost of coding rate loss.

SLM In SLM method, an input data vector of N symbols is multiplied by U phase sequences $\mathbf{P}^u = [P_0^u, P_1^u, \dots, P_N^u]$, $1 \leq u \leq U$ to generate U alternative symbol sequences that represents the same information as the original data block. Inverse discrete Fourier transform (IDFT) is performed for each of the U alternative symbol sequences, and then the one with the lowest PAPR is selected for transmission (Bäumel et al. 1996). In order to recover the original input data vector, information about the selected phase sequence has to be transmitted to the receiver as side information. Also, its computational complexity is quite high due to the operation of U IDFTs.

PTS In PTS method, an input data vector of N symbols is partitioned into M disjoint subblocks. The subcarriers in each subblock are weighted by a phase factor designated for that subblock. The M phase factors are selected such that the PAPR of the combined OFDM signal is minimized (Müller and Huber 1997). Similar to the SLM method, the side information, i.e., the selected phase factors, must be transmitted to the receiver to recover the original data vector. Also, the search complexity of an adequate set of phase factors increases exponentially with the number of subblocks.

Interleaving The interleaving method is similar to the SLM scheme. V interleavers that per-

mute or reorder a block of symbols in respective ways are applied on the input data vector to produce V different permuted data vectors. IDFT is performed on each one of the V permutations separately, and the one with the lowest PAPR is chosen for transmission (Eetvelt et al. 1996). To recover the original data vector, the receiver needs to be informed which interleaver is used at the transmitter. Its computational complexity is also high due to the calculation of V IDFTs.

TR The TR method is to reserve a small set of subcarriers for transmitting PAPR reduction signals instead of data transmission. Let $\mathcal{R}_c$ denote the set of PAPR reduction subcarriers and $\mathcal{R}_d$ denote the set of data subcarriers. The input data vector following the TR method turns to

$$\hat{X}(k) = X(k) + C(k) = \begin{cases} C(k), & k \in \mathcal{R}_c \\ X(k), & k \in \mathcal{R}_d \end{cases}, \quad (6)$$

where $C(k)$ denote the PAPR reduction signals, which are generated to minimize the maximum peak value of $\text{IDFT}(\hat{X}(k))$ through a convex optimization. In TR scheme, the receiver only needs to know the position of the PAPR reduction subcarriers but not the actual PAPR reduction signals $C(k)$. However, the subcarriers reserved for the PAPR reduction cause a data rate decrease. Moreover, additional processing power is required at the transmitter to generate the PAPR reduction signals.

TI The basic idea of TI method is to increase the constellation size of symbols at OFDM subcarriers, so that each of the points in the original constellation can be mapped into several equivalent points in the expanded constellation, prior to the IDFT processing. The extra degrees of freedom, i.e., each symbol can be mapped into one of the several equivalent constellation points, are exploited for PAPR reduction. For a square QAM constellation of size M and minimum distance between constellation points d , the QAM symbol $X(k)$ can be mapped to the expanded constellation as

$$\hat{X}(k) = X(k) + p_k D + q_k D, \quad (7)$$

where $D = \rho d \sqrt{M}$, $\rho \geq 1$ represents the spacing between the original point in the original constellation and its equivalent points in the expanded constellation. p_k and q_k are integer numbers used to change the real and imaginary parts of $X(k)$, which are selected to reduce the PAPR. The amount of PAPR reduction depends on the value of ρ and the number of expanded symbols in a data block. It is noteworthy that the injected signals in the TI method have an increased energy, which would increase the average transmit power.

Hybrid Techniques

Hybrid techniques combine two or more methods for PAPR reduction. The hybridization can have the advantages of the combined methods, albeit with slight increases in complexity. A few studies have validated the combination of coding-based PAPR reduction method with clipping, SLM, and PTS schemes. These combined methods can achieve significant PAPR reduction with a small data rate loss, while the coding technique presents good error control properties (Sandoval et al. 2017).

Key Applications

PAPR reduction is fundamental in the implementation of OFDM systems, particularly when the number of subcarriers is large.

Cross-References

- [Principle of OFDM and Multi-carrier Modulations](#)

References

- Bäumel R, Fischer RFH, Huber J (1996) Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping. IEE Electron Lett 32:2056–2057

- Boyd S (1986) Multitone signals with low crest factor. *IEEE Trans Circuits Syst* 33:1018–1022
- Eetvelt P, Wade G, Tomlinson M (1996) Peak to average power reduction for OFDM schemes by selective scrambling. *IEE Electron Lett* 32: 1963–1964
- Greenstein L, Fitzgerald P (1981) Phasing multitone signals to minimize peak factors. *IEEE Trans Commun* 29:1072–1074
- Jones A, Wilkinson T, Barton S (1994) Block coding scheme for reduction of peak-to-average envelope power ratio of multicarrier transmission systems. *IEE Electron Lett* 30:2098–2099
- Li X, Cimini L (1998) Effect of clipping and filtering on the performance of OFDM. *IEEE Commun Lett* 2:131–133
- Müller S, Huber J (1997) OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences. *IEE Electron Lett* 33: 368–369
- O'Neill R, Lopes L (1995) Envelope variations and spectral splatter in clipped multicarrier signals. In: *IEEE PIMRC95*, pp 71–75
- Rahmatallah Y, Mohan S (2013) Peak-to-average power ratio reduction in OFDM systems: a survey and taxonomy. *IEEE Commun Surv Tuts* 15:1567–1592
- Sandoval F, Poitau G, Gagnon F (2017) Hybrid peak-to-average power ratio reduction techniques: review and performance comparison. *IEEE Access* 5:27145–27161
- Schroeder M (1970) Synthesis of low-peak-factor signals and binary sequences with low autocorrelation. *IEEE Trans Inform Theory* 16:85–89
- Tellado J (1999) Peak to average power ratio reduction for multicarrier modulation. Ph.D. thesis, University of Stanford, California
- van Nee R (1996) OFDM codes for peak-to-average power reduction and error correction. In: *IEEE global telecommunications conference*, pp 740–744
- van Nee R, de Wild A (1998) Reducing the peak-to-average power ratio of OFDM. In: *IEEE vehicular technology conference*, pp 2072–2076
- Wang X, Tjhung TT, Ng CS (1999) Reduction of peak-to-average power ratio of OFDM system using a companding technique. *IEEE Trans Broadcast* 45: 303–307

Parameter Estimation

- [Parameter Estimation Exploiting WSNs with Arbitrary Topology Geometry](#)

Parameter Estimation Exploiting WSNs with Arbitrary Topology Geometry

Liangtian Wan

Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian, China

Synonyms

[Channel estimation](#); [Parameter estimation](#)

Definitions

The deployment of the WSN can form arbitrary topology geometry. The parameter estimation can be estimated based on the received data of the sensors in WSN.

Historical Background

Wireless sensor network (WSN) is a distributed sensing network with hundreds or thousands of sensors that sense and examine the outside world. The sensors in the WSN communicate in a wireless way, so the network settings are flexible. The location of sensors can be changed at any time, and the Internet can be connected with the WSN in a wired or wireless manner. A multi-hop self-organizing network formed by wireless communication. The WSN can be deployed in different areas to monitor the targets of interest (Wan et al. 2015). The topology of the WSN can be arbitrary. When the sensors of WSN are mounted on the surface of complex carrier, they construct a conformal array (Josefsson and Persson 2006). Compared with traditional array, the conformal array has some characteristics. First, due to the irregular structure of the carrier, the sensors of WSN conforming to the surface of the carrier have directionality, i.e., it has different pattern functions and has different responses to incident signals of different directions. Second,

the curvature of the complex carrier surface generally has a radius, and the maximum radiation direction of each sensor is almost different. Third, when the signal is impinging on the surface of the carrier from certain directions, some of the sensors cannot receive the signal, or the received signal is weak due to the occlusion effect of the carrier. These characteristics mentioned above have brought unprecedented challenges to the parameter estimation of WSN. Thus, the parameter estimation for targets of interest based on WSN is an important problem to solve due to the arbitrary topology geometry of the WSN.

The Signal Model with Arbitrary Topology Geometry

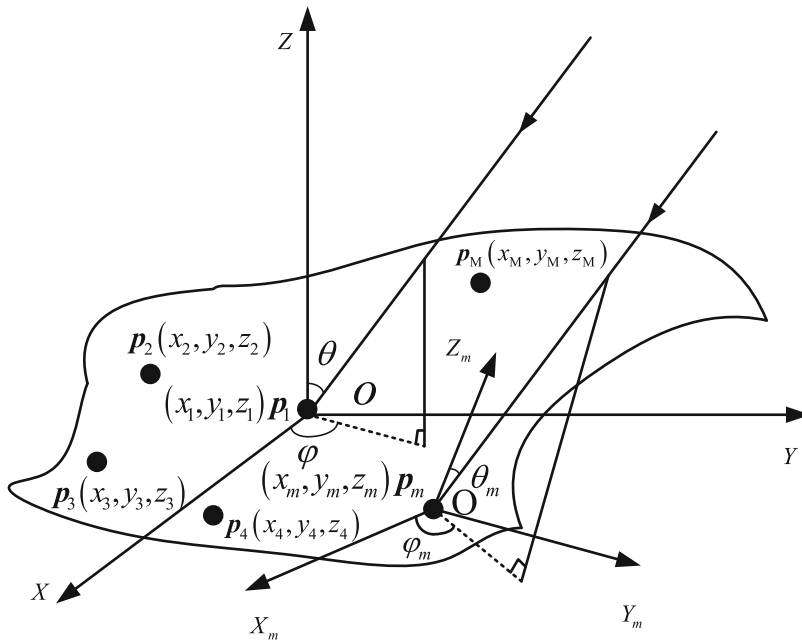
As shown in Fig. 1, a WSN is mounted on the surface of an arbitrarily shaped carrier. Assume that there are M sensors on the surface of carrier. In global coordinate system, the coordinate of each sensor can be expressed as (x_m, y_m, z_m) , $m = 1, 2, \dots, M$. The origin of the global coordinate

system is the position of sensor $\mathbf{p}_1$, and the origins of the local coordinate systems are the positions of sensor $\mathbf{p}_m$. For the identical signal, the azimuth angle and elevation angle are different in global and local coordinate systems. This is different with the traditional array, e.g., uniform linear array (ULA). For the traditional array, there is only one azimuth angle and elevation angle corresponding to the identical signal.

The signals impinging on the arbitrary array consisting of M sensors are narrowband farfield signals, the direction vector $\mathbf{u}$ with frequency f can be expressed as $\mathbf{u} = [-\sin(\theta)\cos(\varphi) \quad -\sin(\theta)\sin(\varphi) \quad -\cos(\theta)]^T$, where θ and φ stand for the elevation angle and azimuth angle, respectively. “T” stands for the transpose of a vector or a matrix.

The Signal Model with Conformal Array

The array manifold of conformal arrays is modeled by geometric algebra in Zou et al.



Parameter Estimation Exploiting WSNs with Arbitrary Topology Geometry, Fig. 1 The structure of the conformal array constructed by WSN

(2011), but the geometric algebra method is too abstract when analyzing the coordinate system rotation, which makes it difficult to establish an intuitive impression of this method. In this paper, the Euler rotation transformation method proposed in Wang et al. (2008) and Wan et al. (2014) is used to realize the conversion between the global coordinate system and the local coordinate system of the sensor.

Consider r signals from different directions (θ_i, φ_i) , $i = 1, 2, \dots, r$ impinge on the conformal array consisting of M sensors, where θ_i and φ_i are the elevation angle and azimuth angle of the i th incident signal in global coordinate system, respectively. The snapshot data matrix received by the conformal array can be expressed as $\mathbf{X}(t) = \mathbf{A}(\theta, \varphi)\mathbf{S}(t) + \mathbf{N}(t)$, where $\mathbf{X}(t)$ is the $M \times 1$ snapshot data vector, $\mathbf{S}(t)$ is the $r \times 1$ signal vector, $\mathbf{N}(t)$ is $M \times 1$ noise vector, and $\mathbf{A}(\theta, \varphi)$ is $M \times r$ manifold matrix of conformal array, which can be expressed as $\mathbf{A}(\theta, \varphi) = [\mathbf{a}(\theta_1, \varphi_1), \mathbf{a}(\theta_2, \varphi_2), \dots, \mathbf{a}(\theta_r, \varphi_r)]$, where $\mathbf{a}(\theta_i, \varphi_i)$ is the steering vector of the i th signal expressed as $\mathbf{a}(\theta_i, \varphi_i) = [g_1(\theta_i, \varphi_i) \exp(-j2\pi \frac{\mathbf{p}_1 \cdot \mathbf{u}}{\lambda}), g_2(\theta_i, \varphi_i) \exp(-j2\pi \frac{\mathbf{p}_2 \cdot \mathbf{u}}{\lambda}), \dots, g_M(\theta_i, \varphi_i) \exp(-j2\pi \frac{\mathbf{p}_M \cdot \mathbf{u}}{\lambda})]^T$, $\mathbf{p}_m = (x_m, y_m, z_m)$, $m = 1, 2, \dots, M$, where $g_m(\theta_i, \varphi_i)$ stands for the direction pattern of the m th sensor for the direction (θ_i, φ_i) in the global coordinate system, “ $\cdot$ ” stand for the Hadamard product, $\mathbf{p}_m$ stands for the position vector of the m th sensor. For the traditional linear array, the sensors in the WSN have the identical direction pattern. Thus $g_m(\theta_i, \varphi_i)$ is normalized as 1. However, for the conformal array, the sensors in the WSN have different direction patterns due to the curvature of the complex carrier surface. The different direction patterns bring new challenges for parameter estimation. The accurate modelling of the direction pattern of $g_m(\theta_i, \varphi_i)$ should be taken into consideration, and the direction pattern of the sensor should be designed in its local coordinate system. Thus a key problem for the modelling of the steering vector of the conformal array is the direction pattern transformation between the local coordinate system and the global coordinate system.

For the global coordinate system and the m local coordinate system, we can, respectively, obtain $x = \rho \sin \theta \cos \varphi$, $y = \rho \sin \theta \sin \varphi$, $z = \rho \cos \theta$ and $x_m = \rho_m \sin \theta_m \cos \varphi_m$, $y_m = \rho_m \sin \theta_m \sin \varphi_m$, $z_m = \rho_m \cos \theta_m$.

We only know the directions of the impinging signals, and we set $\rho = \rho_m = 1$, where θ_m and φ_m stand for the elevation angle and azimuth angle in the m th local coordinate system of the impinging signal.

Then the conversion of each component from the global Cartesian coordinate system to the m th local Cartesian coordinate system should be completed based on three Euler rotation transformations. Each transformation corresponds to a Euler rotation matrix, i.e., the conversion from the global Cartesian coordinate system to the m th local Cartesian coordinate system can be realized by using three orthogonal transformation matrices, $\mathbf{R}_X$, $\mathbf{R}_Y$, and $\mathbf{R}_Z$, where $\mathbf{R}_X$ is expressed as $R_X = [1, 0, 0; 0, \cos \alpha_X, -\sin \alpha_X; 0, \sin \alpha_X, \cos \alpha_X]$. It indicates that the X axis of the global coordinate system is the rotation axis, $-\pi \leq \alpha_X \leq \pi$. Clockwise is positive during the rotation and negative counterclockwise.

$\mathbf{R}_Y$ is expressed as $R_Y = [1, 0, 0; 0, \cos \alpha_Y, -\sin \alpha_Y; 0, \sin \alpha_Y, \cos \alpha_Y]$. It indicates that the Y axis of the global coordinate system is the rotation axis, $-\pi \leq \alpha_Y \leq \pi$. Based on the right-handed spiral rule, we face the negative direction of the X axis and rotate on the YOZ plane. The rotation angle α_X in the equation can be expressed as the angle between the Y axis (or Z axis) of the new coordinate system and the Y axis (or Z axis) of the original coordinate system.

$\mathbf{R}_Z$ is expressed as $R_Z = [1, 0, 0; 0, \cos \alpha_Z, -\sin \alpha_Z; 0, \sin \alpha_Z, \cos \alpha_Z]$. It indicates that the Z axis of the global coordinate system is the rotation axis, $-\pi \leq \alpha_Z \leq \pi$. Based on the right-handed spiral rule, we face the negative direction of the Z axis and rotate on the XOY plane. The rotation angle α_Z in the equation can be expressed as the angle between the X axis (or Y axis) of the new coordinate system and the X axis (or Y axis) of the original coordinate system.

Therefore, the conversion relationship between the global Cartesian coordinate system and the m th local Cartesian coordinate system can be obtained $[x_m; y_m; z_m] = \mathbf{R}_X \mathbf{R}_Y \mathbf{R}_Z [x; y; z]$.

The design of the pattern of the sensor is usually based on the local coordinate system of the sensor itself. Therefore, it is necessary to obtain the elevation angle and azimuth angle of the impinging signal in the local coordinate system. The conversion between the spherical coordinate system and the Cartesian coordinate system can be expressed as $\theta_m = \arccos z_m$, $\varphi_m = \arctan(\frac{y_m}{x_m})$. The transform between global Cartesian coordinate system and the local Cartesian coordinate system is achieved.

Conclusions

In this entry, we introduce a coordinate rotation transformation method for conformal array consisting of sensors in the WSN. Based on this method, the WSN with arbitrary topology configuration can be used for the parameter estimation, e.g., direction of arrival, etc.

Future Directions

In the future, we will design novel model for conformal array and develop parameter estimation algorithms for WSN with arbitrary.

Key Application

The parameter estimation with arbitrary topology geometry can be used for detecting and tracking the moving targets. In addition, the conformal array can be mounted on the aircraft to reduce the air resistance. This structure has the smallest radar cross section, and the aircraft is hard to detected.

Cross-References

- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)
- [Localization in 3D Surface Wireless Sensor Networks](#)

References

- Josefsson L, Persson P (2006) Conformal array antenna theory and design, vol 29. Wiley, Hoboken
- Wan L, Si W, Liu L, Tian Z, Feng N (2014) High accuracy 2D-DOA estimation for conformal array using PARAFAC. *Int J Antennas Propag* 2014:1–14
- Wan L, Han G, Shu L, Feng N, Zhu C, Lloret J (2015) Distributed parameter estimation for mobile wireless sensor network based on cloud computing in battlefield surveillance system. *IEEE Acc* 3:1729–1739
- Wang BH, Guo Y, Wang YL, Lin YZ (2008) Frequency-invariant pattern synthesis of conformal array antenna with low cross-polarisation. *IET Microw Antennas Propag* 2(5):442–450
- Zou L, Lasenby J, He Z (2011) Beamforming with distortionless co-polarisation for conformal arrays based on geometric algebra. *IET Radar Sonar Navig* 5(8):842–853

Recommended Reading

- Josefsson L, Persson P (2006) Conformal array antenna theory and design. Wiley Online Books, Piscataway
- Van Trees HL (2002) Optimum array processing: part IV of detection, estimation, and modulation theory. Wiley Online Books, New York

Pareto Efficiency

- [Efficiency and Pareto Optimality](#)

Particle-Based Simulation

- [Reaction-Diffusion Channels](#)

Path Loss and Interference Analysis for a Random Two-Hop Relay Link

Ruonan Zhang¹ and Jianping Pan²

¹Northwestern Polytechnical University, Shaanxi, China

²University of Victoria, Victoria, BC, Canada

Technical Background

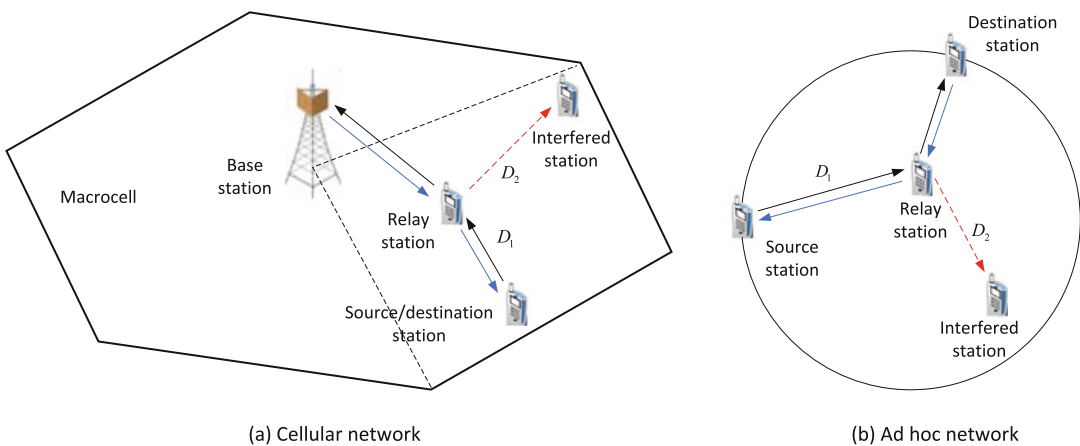
In relay communications, signals from a source station (SS) are relayed to a destination station (DS) by one or more relay stations (RSs), as shown in Fig. 1. In radio propagation, signal power decays with the increase of transmission distance superlinearly with a certain exponent (3GPP 2017; Wang et al. 2017). By reducing the propagation distance in each hop through a relay link, the total transmission power is reduced, and each hop can achieve a higher data rate. Consequently, the relay communication leads to a higher spectrum and power efficiency and end-to-end throughput. Meanwhile, the coverage range of a station is increased significantly. The cooperative communications in cellular networks and the peer-to-peer communications in ad hoc wireless networks are both based on the relay communications and have attracted considerable

research effort in recent years (Sheng et al. 2011; Zhang et al. 2016; Qin et al. 2015) (Table 1).

A relay link also generates interference to other stations in a wireless network, as illustrated in Fig. 1. The interference is herein the signals transmitted from an RS and received by a non-destination station. The station under interference is referred to as an *interfered station (IS)*. Suppose that the RSs employ the amplify-and-forward (AF) mode by which each RS amplifies the received radio signals and directly transmits them without decoding and encoding. The AF scheme is easy to implement in practice when compared to the others such as the decode and forward (DF) and cooperative coding (CC) schemes. Therefore, the interference for AF relay scheme is discussed in this entry.

According to Fig. 1, let $P_{t,s}$ and $P_{t,h}$ denote the transmission power at the SS and RS (helper), respectively, and $P_{r,h}$ and $P_{r,i}$ denote the received power at the RS and IS, respectively. Let G_s , G_h , and G_i be the antenna gains at the SS, RS, and IS, respectively. Suppose that the distances from the SS to the RS and from the RS to the IS are $D_{s,h}$ and $D_{h,i}$, respectively. Based on the close-in (CI) free space reference distance model, the path loss (PL) of the first hop (from the SS to the RS), denoted by $L_{s,h}$, is

$$L_{s,h} = \frac{P_{t,s}}{P_{r,h}} = \frac{1}{G_s G_h} \left(\frac{4\pi d_0}{\lambda} \right)^2 \left(\frac{D_{s,h}}{d_0} \right)^\alpha, \quad (1)$$



Path Loss and Interference Analysis for a Random Two-Hop Relay Link, Fig. 1 Two-hop relay communications and interference in a wireless network. (a) Cellular network. (b) Ad hoc network

Path Loss and Interference Analysis for a Random Two-Hop Relay Link, Table 1 Definitions in this article

Interference	Addition of unwanted signals which modify or disrupt a useful signal
Source station	A station that generates and sends original information
Destination station	A station designated as the receiver of information
Relay station	A station that receives and passes along signals or information to another station
Interfered station	A station that is interfered by the signals from other stations
Cellular network	A wireless communication network where the access service is organized in cells
Ad hoc network	A decentralized wireless network where each node may forward data for other nodes
Path loss	Attenuation of power density when a radio signal propagates over a distance
Antenna gain	Amplification of signal power by an antenna at a specific direction

where λ is the signal wavelength and α is the path loss exponent (PLE). The d_0 in (1) is the reference distance and usually set to 1 ~ 10 meters for indoor and 1 kilometer for outdoor scenarios. The PL between the RS and IS, denoted by $L_{h,i}$, is defined similarly by (1) with the antenna gains and propagation distance between the RS and IS accordingly.

The average received power is mainly determined by the PL of each hop, with shadowing and fading which are often distance independent. Suppose that the amplification gain of the RS is A_{rel} in the AF relaying. The interference power at the IS is

$$P_{r,i} = \frac{P_{t,h}}{L_{h,i}} = \frac{A_{rel} P_{t,s}}{L_{s,h} L_{h,i}} = \beta (D_{s,h} D_{h,i})^{-\alpha}, \quad (2)$$

where

$$\beta = A_{rel} G_s G_h^2 G_i P_{t,s} d_0^{2\alpha} \left(\frac{\lambda}{4\pi d_0} \right)^4. \quad (3)$$

The system parameters in (2), such as the transmission power, amplification gain, and antenna gains, are usually fixed by the hardware design and implementation. Thus, the term β is a constant and the interference power at the IS, $P_{r,i}$, is proportional to the term $(D_{s,h} D_{h,i})^{-\alpha}$. Therefore, the distribution of the product of the two transmission distances, $D_{s,h} D_{h,i}$, determines the statistics of the interference over the two-hop link. However, the RS is usually randomly located in a wireless network, depending on the available resources, signal relay efficiency, and motivation to help other stations (Zhuang and Zhou 2013). Thus the

distances $D_{s,h}$ and $D_{h,i}$ are stochastic and dependent. Let $F_{D_{s,h} D_{h,i}}(d)$ and $f_{D_{s,h} D_{h,i}}(d)$ denote the CDF and PDF of the distance product, respectively. The CDF of the interference power is

$$\begin{aligned} F_{P_{r,i}}(p) &= \Pr \{ P_{r,i} < p \} = \Pr \{ \beta (D_{s,h} D_{h,i})^{-\alpha} < p \} \\ &= \Pr \left\{ (D_{s,h} D_{h,i}) > \left(\frac{\beta}{p} \right)^{\frac{1}{\alpha}} \right\} \\ &= 1 - F_{D_{s,h} D_{h,i}} \left[\left(\frac{\beta}{p} \right)^{\frac{1}{\alpha}} \right]. \end{aligned} \quad (4)$$

By taking the derivative of $F_{P_{r,i}}(p)$ with respect to p , the PDF of the interference power can be obtained as

$$f_{P_{r,i}}(p) = \frac{dF_{P_{r,i}}(p)}{dp} = \frac{\beta^{\frac{1}{\alpha}}}{\alpha p^{\frac{1}{\alpha}+1}} f_{D_{s,h} D_{h,i}}(d). \quad (5)$$

With the distributions of the received signal and interference strength, signal-to-interference-plus-noise ratio (SINR), link capacity, and outage probability can be obtained accordingly. Please note that although the CI model is adopted to derive the interference power from (1) to (3) here, other PL models can be used, such as the floating-intercept (FI) model. Because the PL is generally increased superlinearly with the distance powered to the PLE, the interference power is proportional to $(D_{s,h} D_{h,i})^{-\alpha}$. Thus the results in (4) and (5) and the approach are applicable to

other PL models, and the coefficient β is changed accordingly.

Key Applications

The interference model for a random two-hop relay link above can be used to analyze and manage interference in a wireless network. As shown in (5), the key issue is the distribution functions of the distance product in the relay link, considering the random location of the RS. The stochastic geometry approach can be employed to derive $F_{D_{s,h}D_{h,i}}(d)$ and $f_{D_{s,h}D_{h,i}}(d)$ in a random two-hop relay link (Song et al. 2013; Baccelli and Giovanidis 2015; Zhuang et al. 2010). For example, the following scenarios can be studied:

- For a cellular network, the BS communicates with a target user equipment (UE) via a two-hop relay link in a hexagonal cell. The BS is located at the center, and the RS is uniformly located inside the rhombus area covered by a directional antenna of the BS. The IS may be in the same or an adjacent cell, and it is interfered by the signal transmitted by the RS.
- For an ad hoc network, two stations communicate via a two-hop link without the constraint of their locations. The RS is uniformly located in a lens-like area determined by the SS and the DS.

Song et al. (2013) obtained the closed-form and approximation expressions of the distribution functions of the distance product, and similar approach can be extended to the other scenarios.

By plugging $F_{D_{s,r}D_{r,i}}(d)$ and $f_{D_{s,r}D_{r,i}}(d)$ into (4) and (5), the distribution functions of the interference power can be obtained. Then the network performance and interference management algorithms can be evaluated analytically with more insights on their properties.

References

3rd Generation Partnership Project (2017) Study on channel model for frequencies from 0.5 to 100 GHz (Release 14). 3GPP Technical Specification Group Radio Access Networks, Tech. Rep. TR38.901 V14.0.0

- Baccelli F, Giovanidis A (2015) A stochastic geometry framework for analyzing pairwise-cooperative cellular networks. *IEEE Trans Wireless Commun* 14(2): 794–808
- Qin H, Zhang R, Li B, Cai L (2015) Distributed cooperative MAC for wireless networks based on network coding. In: *Proceedings of the IEEE wireless communications and networking conference (WCNC)*, New Orleans, pp 2050–2055
- Sheng Z, Leung KK, Ding Z (2011) Cooperative wireless networks: from radio to network protocol designs. *IEEE Commun Mag* 49(5):64–69
- Song X, Zhang R, Pan J, Liu J (2013) A statistical geometric approach for capacity analysis in two-hop relay communications. In: *Proceedings of the IEEE global communications conference (GLOBECOM)*, Atlanta, pp 4823–4829
- Wang K, Zhang R, Wu L, Zhong Z, He L, Liu J, Pang X (2017) Path loss measurement and modeling for low-altitude UAV access channels. In: *Proceedings of the IEEE vehicular technology conference (VTC 2017-Fall)*, Toronto, pp 1–5
- Zhang R, Ban D, Li B, Jiang Y (2016) The cooperative multicasting based on random network coding in wireless networks. In: *Proceedings of the IEEE vehicular technology conference (VTC 2016-Spring)*, Nanjing, Jiangsu, China, pp 1–5
- Zhuang W, Zhou Y (2013) A survey of cooperative MAC protocols for mobile communication networks. *J Internet Technol* 14(4):541–559
- Zhuang Y, Pan J, Cai L (2010) Minimizing energy consumption with probabilistic distance models in wireless sensor networks. In: *Proceedings of the IEEE international conference on computer communications (INFOCOM)*, San Diego, pp 1–9

PD-NOMA

- [Non-orthogonal Multiple Access](#)

Peak Factor Reduction

- [PAPR Reduction in OFDM Systems](#)

Peer-to-Peer Energy Trading of Microgrids

- [Intelligence and Trading with Energy Storage and Renewable Generation of Microgrids](#)

Per-Beam Synchronization

► Per-Beam Synchronization for Millimeter-Wave Massive MIMO

Per-Beam Synchronization for Millimeter-Wave Massive MIMO

Li You¹, Xiqi Gao¹, Geoffrey Ye Li², Xiang-Gen Xia³, and Ni Ma⁴

¹National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

²School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

³Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

⁴Huawei Technologies Co., Ltd., Shenzhen, China

Synonyms

Massive MIMO; Millimeter-wave communications; Per-beam synchronization

Definitions

Per-beam synchronization refers to a synchronization approach in wireless transmission where signal synchronization in time and frequency between the transmitter and the receiver is performed over each beam individually.

Background

Millimeter-wave massive multiple-input multiple-output (MIMO) is promising for future wireless communications (Swindlehurst et al. 2014; Cui et al. 2018; Huang et al. To appear; Liu

et al. 2018; Xiao et al. 2017). One challenge in realizing cellular wireless over millimeter-wave channels is to deal with the mobility issue (Roh et al. 2014; Andrews et al 2017). For the same mobile speed, the Doppler spread of millimeter-wave channels is orders-of-magnitude larger than that of classical wireless channels, while the delay spread does not change significantly over different frequencies, which may lead to system implementation bottleneck. Consider wideband millimeter-wave transmission employing orthogonal frequency division multiplexing (OFDM) modulation, for example. With perfect time and frequency synchronization in the space domain, the length of the cyclic prefix (CP) is usually set to be slightly larger than the delay span to mitigate channel dispersion in time, while the length of the OFDM symbol is usually set to be inversely proportional to the Doppler spread to mitigate channel dispersion in frequency (Hwang et al. 2009; Dahlman et al. 2014). As a result, the overhead of the CP will be much larger to deal with the same delay spread, and it might be difficult to design proper OFDM parameters.

There exist some works related to the above issue. For example, beam-based Doppler frequency compensation has been suggested in Chizhik (2004) and Va et al. (2017) for narrowband MIMO channels. In addition, reduced delay spread with narrow directional beams has been observed in recent millimeter-wave channel measurement results (Rappaport et al. 2015).

Massive MIMO Beam Domain Channel Model

Consider a single-cell massive MIMO system, where the base station equipped with an M -element uniform linear array (ULA) serves several users, each with a K -element ULA. Due to the small wavelength in millimeter-wave bands, the case where both the numbers of antennas at the base station and the users are sufficiently large is of practical interest (Andrews et al 2017), which is different from the massive

MIMO communications over lower frequency bands (Marzetta 2010).

Denote the complex baseband downlink of space domain channel frequency response matrix of user u as $\mathbf{G}_u$. Then the corresponding beam domain channel matrix can be represented as $\bar{\mathbf{G}}_u = \mathbf{F}_K^H \mathbf{G}_u \mathbf{F}_M^*$ where $\mathbf{F}_K$ is the $K \times K$ unitary discrete Fourier transform (DFT) matrix. Both transformation matrices, $\mathbf{F}_K$ and $\mathbf{F}_M$, can be interpreted as DFT beamforming operations performed at the base station and the users, respectively. Thus, $\bar{\mathbf{G}}_u$ is referred to as the beam domain channel frequency response matrix.

In the asymptotically large array regime, the fading of each of the beam domain channel elements tends to disappear when both the numbers of antennas at the base station and the users tend to infinity. The physical interpretation is intuitive. Specifically, beamforming can effectively divide the channels in the angle domain, and the partition resolution can be sufficiently high with sufficiently large numbers of antennas at both the base station and the users. Asymptotically, each propagation path can be resolved, and the beam domain channel element corresponds to the gain of a specific propagation path along a fixed AoA-AoD pair. Thus, the beam domain channel envelopes tend to remain a constant in both the time and the frequency domains (You et al. 2017).

Per-Beam Synchronization in Time and Frequency

Consider the OFDM-based millimeter-wave massive MIMO wireless transmission. The performance of OFDM-based transmission is sensitive to time and frequency offsets, so it is necessary to perform time and frequency synchronization to compensate for time and frequency offsets of the received signals. In particular, the received signals should be carefully adjusted so that the resultant minimum time offset and center frequency offset are aligned to zero (Hwang et al. 2009). The most common synchronization approach for MIMO systems is to compensate for the time and frequency offsets of the received signals in the space domain using the same time

and frequency adjustment parameters. Then the resultant channel frequency spread scales linearly with the carrier frequency for a given mobile velocity. Therefore, in order to support the same user mobility, the length of the OFDM symbol in millimeter-wave systems would be substantially reduced compared with that in the conventional microwave systems. Meanwhile, the length of the CP would be the same as that in the conventional microwave systems to deal with the same delay spread, which might lead to difficulty in selecting proper OFDM parameters.

Motivated by the beam domain massive MIMO channel properties in the time and frequency domains, per-beam synchronization in time and frequency, where adjustment of time and frequency offsets is applied to the signal over each receive beam individually, can be adopted to address the above issue (You et al. 2017). Compared with the conventional synchronization approach, the effective channel delay and frequency spreads can be reduced with the per-beam synchronization approach. In particular, the effective channel frequency spread is approximately reduced by a factor of the number of user antennas in the large array regime. In addition, the effective channel delay spread can also be reduced with per-beam synchronization, but the quantitative performance is difficult to establish. For the case in which a ring of scatterers is located around the users (Shiu et al. 2000; Pätzold and Hogstad 2006), the effective channel delay spread with per-beam synchronization is also reduced by a factor of the number of user antennas.

Per-beam synchronization can be exploited to simplify the implementation and improve the performance of millimeter-wave massive MIMO-OFDM systems. In particular, the number of user antennas also scales linearly with the carrier frequency for the same antenna array aperture although the maximum channel Doppler shift scales linearly with the carrier frequency. Thus, assuming a fixed antenna array aperture, the effective channel Doppler frequency spread over millimeter-wave bands becomes approximately the same as that over regular bands with per-beam synchronization, which can mitigate severe

Doppler effects over millimeter-wave channels. Moreover, the effective channel delay spread can be significantly reduced with per-beam synchronization, which can further lead to a substantial reduction in the CP overhead.

Key Applications

Per-beam synchronization can be adopted to deal with the mobility issue in millimeter-wave massive MIMO. The motivation of per-beam synchronization stems from the beam domain massive MIMO channel property that the fading of each of the beam domain channel elements tends to disappear in the asymptotic large array regime. For per-beam synchronization, signal synchronization in time and frequency between the transmitter and the receiver is performed over each beam individually. With per-beam synchronization, both the effective channel delay and Doppler frequency spreads can be approximately reduced by a factor of the number of antennas at the receiver in the large array regime with per-beam synchronization compared with the conventional synchronization approaches, which effectively mitigates the severe Doppler effect in millimeter-wave systems and leads to a significantly reduced CP overhead. Per-beam synchronization can be applied in massive MIMO transmission approaches such as the beam division multiple access transmission (Sun et al. 2015).

Cross-References

► [Millimeter Wave Massive MIMO](#)

References

- Andrews JG, Bai T, Kulkarni MN, Alkhateeb A, Gupta AK, Heath RW Jr (2017) Modeling and analyzing millimeter wave cellular systems. *IEEE Trans Commun* 65(1):403–430
- Chizhik D (2004) Slowing the time-fluctuating MIMO channel by beam forming. *IEEE Trans Wireless Commun* 3(5):1554–1565
- Cui Y, Fang X, Fang Y, Xiao M (2018) Optimal non-uniform steady mmWave beamforming for high speed railway. *IEEE Trans Veh Technol* 67(5):4350–4359. <https://doi.org/10.1109/TVT.2018.2796621>
- Dahlman E, Parkvall S, Sköld J (2014) 4G LTE/LTE-advanced for mobile broadband, 2nd edn. Academic Press, Waltham
- Huang Y, Zhang J, Xiao M (To appear) Constant envelope hybrid precoding for directional millimeter-wave communications. *IEEE J Sel Areas Commun (JSAC) Ser Phys Layer Secur 5G Wirel Netw*. <https://doi.org/10.1109/JSAC.2018.2825820>
- Hwang T, Yang C, Wu G, Li S, Li GY (2009) OFDM and its wireless applications: a survey. *IEEE Trans Veh Technol* 58(4):1673–1694
- Liu Y, Fang X, Xiao M, Mumtaz S (2018) Decentralized beam pair selection in multi-beam millimeter-wave networks. *IEEE Trans Commun*. <https://doi.org/10.1109/TCOMM.2018.2800756>
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wireless Commun* 9(11):3590–3600
- Pätzold M, Hogstad BO (2006) A wideband space-time MIMO channel simulator based on the geometrical one-ring model. In: *Proceedings of IEEE VTC Fall, Montreal*, pp 1–6
- Rappaport TS, MacCartney GR Jr, Samimi MK, Sun S (2015) Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design. *IEEE Trans Commun* 63(9):3029–3056
- Roh W, Seol JY, Park J, Lee B, Lee J, Kim Y, Cho J, Cheun K, Aryanfar F (2014) Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results. *IEEE Commun Mag* 52(2):106–113
- Shiu DS, Foschini GJ, Gans MJ, Kahn JM (2000) Fading correlation and its effect on the capacity of multielement antenna systems. *IEEE Trans Commun* 48(3):502–513
- Sun C, Gao XQ, Jin S, Matthaiou M, Ding Z, Xiao C (2015) Beam division multiple access transmission for massive MIMO communications. *IEEE Trans Commun* 63(6):2170–2184
- Swindlehurst AL, Ayanoglu E, Heydari P, Capolino F (2014) Millimeter-wave massive MIMO: the next wireless revolution? *IEEE Commun Mag* 52(9):56–62
- Va V, Choi J, Heath RW Jr (2017) The impact of beamwidth on temporal channel variation in vehicular channels and its implications. *IEEE Trans Veh Technol* 66(6):5014–5029
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis G, Bjornson E, Yang K, Chih-lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun (JSAC)*. 35(9):1909–1935
- You L, Gao XQ, Li GY, Xia XG, Ma N (2017) BDMA for millimeter-wave/Terahertz massive MIMO transmission with per-beam synchronization. *IEEE J Sel Areas Commun* 35(7):1550–1563

Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization

Kalika Suksomboon
Future Communication System Laboratory,
KDDI Research Inc, Fujimino, Saitama, Japan

Synonyms

Configuration and management; Network function virtualization; Packet forwarding function; Performance prediction; Software router

Definitions

Performance modeling of a virtual network function (VNF) refers to a mathematical framework for characterizing the performance of a VNF under the configuration and allocated resources of the hosting commodity hardware.

Historical Background

Wireless network virtualization (WNV) has emerged as a key technology toward a road to the fifth-generation mobile communication (5G) networks. It enables the sharing of radio spectrum and infrastructure resources among multiple 5G services as well as multiple service providers. The adoption of virtualization is for facilitating the flexible and scalable deployment of 5G appliances as virtual network functions (VNFs) running on commodity hardware. Although this operational technique of WNV seems to be reasonable and cost-effective, it requires rethinking since the immaturity of VNF deployment is still a major concern (Mijumbi et al. 2016). This concern translates into the pressing need for the ability to quantify the performance of VNFs running on commodity hardware equivalent to that obtained from network functions (NFs) running on specialized hardware. In other words, sufficient allocated

resources and the optimal configuration of commodity hardware should be specified in order for a VNF to perform at its quality of service (QoS) requirement.

A common VNF for 5G virtual appliances running in the core network is a *packet forwarding function*. The critical challenge disrupting the rapid deployment of a *virtual packet forwarding function* (vPFF) is that different configurations of the hosting server yield different performances in terms of throughput and packet latency, making it impossible to guarantee the QoS. The throughput and packet latency of a vPFF regarding the allocated resources by a precise performance modeling has been seen as a key to quantify the sufficient amount of resources that should be assigned to the vPFF. Moreover, the performance modeling is also beneficial to properly configure a server for serving a vPFF such that its QoS can be guaranteed.

Performance modeling for a packet forwarding function of a hardware router has been extensively studied for decades, e.g., Kim and Kim (2003) and Kiasari et al. (2013). Due to lacks of considering the effect of processor sharing and the interaction of CPU processing with an interrupt handling scheme in a server, these existing analytical models may not be applied for modeling the performance of a vPFF. Alternatively, simulation (Meyer et al. 2014; Beifub et al. 2015) and machine learning (ML) (Riccobene et al. 2016) approaches have been gained more attention as tools for estimating the performance of a vPFF. However, these approaches require a large number of samples of measured data for calibration.

Apart from the aforementioned techniques, a mathematical approach for modeling the performance of a vPFF with a two-dimensional Markov chain has been studied in Gallardo et al. (2016). The model was built upon the assumption that the time the packets wait in Rx queues depends on the implemented polling strategies. However, this model requires the mean packet processing time, mean server vacation time, and queue length as input in order to derive the average packet latency. This approach may not be possible for the realistic WNV systems since the mean packet

processing time and mean server vacation time could not be measured. Hence, this article will present a simple and precise performance modeling of vPFF and introduce an integration of the performance modeling into NFV management and orchestration (NFV-MANO).

Performance Modeling of a Virtual Packet Forwarding Function

Maximum Throughput

The studies in Gallenmüller et al. (2015) and Suk-somboon et al. (2018) reveal that the maximum throughput of a vPFF depends on the allocated resources and configuration of a host server, i.e., Ethernet bandwidth, CPU speed, the number of receiving (Rx) queues, and the number of concurrently running CPUs.

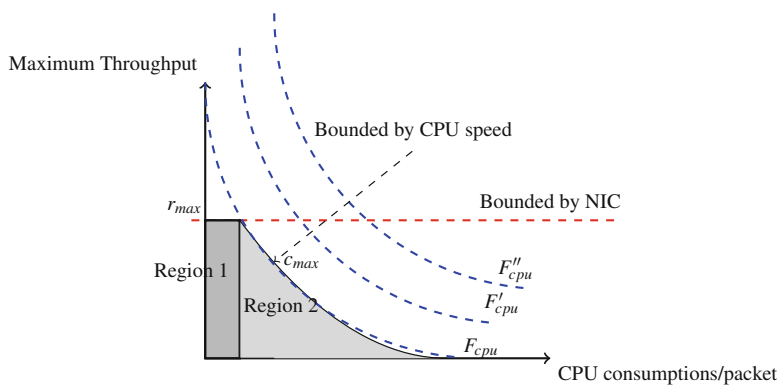
Figure 1 illustrates that the maximum throughput of vPFF is bounded by Ethernet bandwidth and CPU speed. While larger Ethernet bandwidth and higher CPU speeds yield higher maximum throughput, more Rx queues cause the lower maximum throughput. Regarding the observations in Suksoomboon et al. (2018), the average CPU consumption in cycles per packet is linearly scaled by the number of Rx queues, denoted by q . Denote m the maximum number of

Rx queues. Thus, the average CPU consumption per packet can be formulated as a linear function of the number of Rx queues q and q' , where $q \neq q'$ and $q, q' = 1, 2, \dots, m$, as following,

$$C(q) = C(q') + (q - q')\alpha, \quad (1)$$

where $C(q)$ and $C(q')$ are the average CPU consumption per packet with q and q' Rx queues, respectively, and α is the extra CPU consumption due to an added Rx queue. Note that the formula in (1) is flexible enough, thereby observing any two different Rx queue configurations. Hence, the maximum throughput can be computed by minimizing between the maximum throughput bounded by Ethernet bandwidth (i.e., $r_{\max} \leq \frac{v_{\text{eth}}}{b}$ for packet size b and v_{eth} Ethernet bandwidth) and the maximum throughput given q Rx queues bounded by CPU speed F_{cpu} (i.e., $c_{\max}(q) \leq \frac{F_{\text{cpu}}}{C(q)}$).

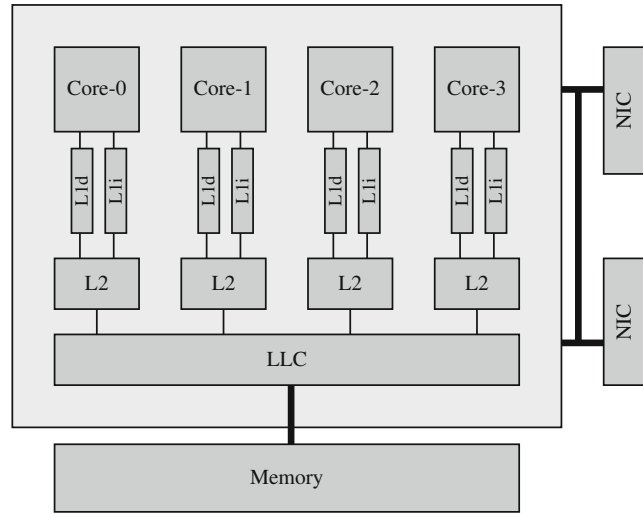
Besides, allocating more the number of CPU cores to vPFF can increase the maximum throughput of a vPFF. However, their relationship is nonlinear due to the cache contention. Figure 2 demonstrates the CPU architecture for analysis. Define a *solo run* as a single CPU core that processes a packet forwarding function without sharing the last-level cache (LLC) and *corunners* as the other CPU cores that concurrently run the



Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization, Fig. 1 Relationship between CPU consumptions per packet and maximum throughput. (Region 1 represents the area where the maximum throughput is bounded by Ethernet

bandwidth, and Region 2 represents the area where the maximum throughput is bounded by CPU speeds F_{cpu} .) Higher CPU speed (e.g., $F_{\text{cpu}} < F''_{\text{cpu}} < F'_{\text{cpu}}$) enlarges Region 2

Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization, Fig. 2 An example of CPU architecture with four CPU cores in a single socket



same function with another CPU core, where two or more CPU cores are assigned to a vPFF and these CPU cores share the same LLC. Regarding the observations in Suksomboon et al. (2017), the CPU consumption per packet is *dilated* from that of a single CPU core due to the cache contention by the corunners, which share the same LLC. The linear increase of the number of cache misses at LLC corresponds to the increasing number of CPU cores assigned to a vPFF. Thus, the *dilated CPU consumption (DCC)* is affected by cache contention in which cache hits turn to misses.

Let $\bar{\alpha}$ be the extra CPU consumption spent for processing a packet once all cache hits of a solo run become cache misses by corunners. Let h be the number of cache hits per second at the LLC with a solo run. Let λ_{solo} be the maximum throughput of a solo run. Therefore, $\frac{h}{\lambda_{\text{solo}}}$ refers to the number of cache hit references per packet at the LLC of a solo run. Hence, the number of cache hit references per packet with corunners becomes $\frac{nh}{\lambda_{\text{solo}}}$, where n is the number of CPU cores. Define ξ as the time difference between accessing the LLC and the system memory. Thus, ξ can be transformed to the number of CPU cycles by ξF_{cpu} cycles. Hence, $\bar{\alpha}$ can be estimated by

$$\bar{\alpha} = \frac{\xi F_{\text{cpu}} n h}{\lambda_{\text{solo}}}. \quad (2)$$

To generalize (2), let $F_{\text{cpu},i}$ be the CPU speed of core i . Hence, $\bar{\alpha} = \frac{\xi h \sum_{j=1}^n F_{\text{cpu},i}}{\lambda_{\text{solo}}}$, and $\bar{\alpha}$ in (2) can be seen as its special form, when all cores have the same CPU speed.

The DCC is estimated upon the assumption that not all but some packets will suffer the cache contention. This assumption is based on the fact that only cache hits turning to cache misses cause CPU spends more cycles to retrieve data from the memory. Meanwhile, the probability of cache misses turning to hits is very small, so this value can be omitted. Therefore, the probability of cache misses being still misses is close to one. Let p be the probability of cache references that were hits at a solo run being still hits with corunners, called *cache-hit-to-hit*—and $p' = 1 - p$ be the probability of cache references that were hits at a solo run becoming misses with corunners—called *cache-hit-to-miss*. Consequently, the packets are divided into two groups: (1) the packets that did not suffer the cache contention and (2) the packets that suffered the cache contention. CPUs spend C cycles for processing packets of the first group and $C + \bar{\alpha}$ for the second group. Let p_i be the probability that CPU core i will not suffer the cache contention by corunning with the other CPU cores j , where $j \neq i$ and $j = 1, \dots, n$, which share the LLC with CPU core i , and $1 - p_i$ is vice versa. This p_i can be referred to as the probability of cache-hit-to-hit,

and $1 - p_i$ can be referred to as the probability of cache-hit-to-miss. Let u_i be the CPU utilization of CPU core i , where $0 \leq u_i \leq 1$ for $i = 1, \dots, n$. The probability of cache-hit-to-hit is obtained by $p = p_i = \min \left\{ \frac{1}{\sum_{j=1}^n u_j}, 1 \right\} \approx \frac{1}{n}$, where $u_j \approx 1$ since a server is overloaded. The maximum throughput for multicore processing systems can be estimated by

$$c_{\max} \approx n F_{\text{cpu}} \left(\frac{C + p\bar{\alpha}}{C(C + \bar{\alpha})} \right). \quad (3)$$

where $F_{\text{cpu},i} = F_{\text{cpu}}$. Note that $\sum_{i=1}^n u_i \geq 1$, where $n > 1$.

Packet Latency

Regarding the observations in Suksomboon et al. (2018), the shapes of the packet latency distributions of vPFF are likely asymmetric bell-shaped distributions, thereby fitting the Erlang- k distributions. The probability density function (PDF) is given by

$$f(x; (k, \beta, \gamma)) = \frac{\beta^k (x - \gamma)^{k-1} e^{-\beta(x-\gamma)}}{(k-1)!}, \quad (4)$$

where the shape parameter $k \in \mathbb{I}^+$ is a positive integer, the scale parameter $\beta \in \mathbb{R}^+$ is a positive real number, and the location parameter $\gamma \in \mathbb{R}^+$ is a positive real number (i.e., $\gamma = \arg \min_x \{x : f(x; (k, \beta, \gamma)) > 0\}$). Fitting an Erlang- k distribution to a packet latency distribution is to select k , β , and γ for minimizing the least-squares error between the measurement result and the fitting curve. The average packet latency can be calculated by the mean of the Erlang- k distribution equal to $\frac{k}{\beta} + \gamma$, and its variance is equal to $\frac{k}{\beta^2}$. Note that the curve fitting method is not for modeling the average packet latency, but it is used for studying the relationship of the Erlang- k parameters to the configured and the observed parameters.

Let $\tau(s)$ be the average packet latency corresponding to a vector of configuration parameters denoted by s , where $s := (q, R, B, F_{\text{cpu}})$ is a four-tuple of q Rx queues, ring size buffer R , netdev_budget B , and CPU speed F_{cpu} . The

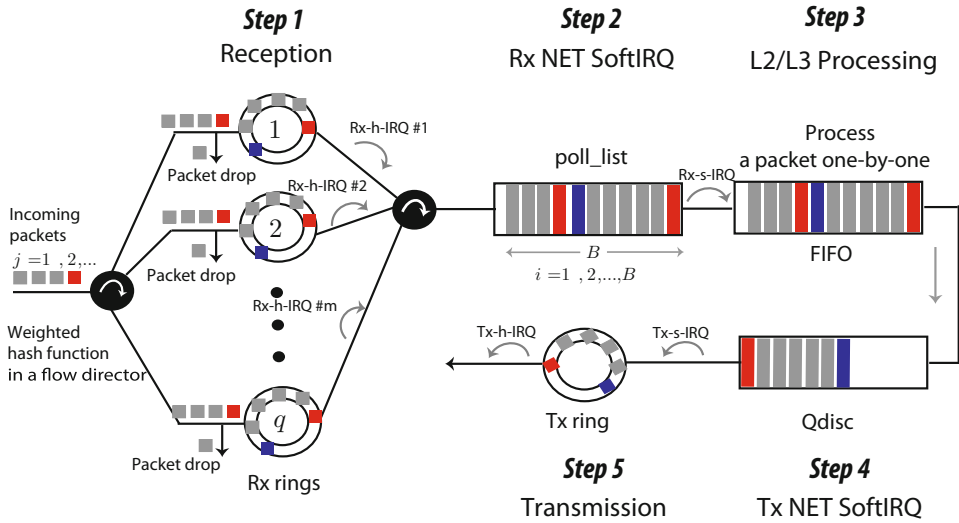
average packet latency can be obtained by estimating $k(s)$, $\beta(s)$, and $\gamma(s)$ as functions of s and substituting them into $f(x; (k(s), \beta(s), \gamma(s)))$ as

$$\tau(s) = E[f(x; (k(s), \beta(s), \gamma(s)))] = \frac{k(s)}{\beta(s)} + \gamma(s). \quad (5)$$

To estimate the Erlang- k parameters, it is required to find the relationship between the Erlang- k parameters to the configured parameters and the measurable parameters of a server's behavior. The observations in Suksomboon et al. (2018) demonstrates the relationship between the Erlang- k parameters to the configured parameters. Nevertheless, this relationship is not sufficient for estimating the Erlang- k parameters. To this end, it requires the understanding of the relationship between the Erlang- k parameters and the measurable parameters of a server's behavior as shown in Fig. 3.

According to the physical meanings of Erlang- k distribution, the packet latency can be represented as a random variable X that has an Erlang- k distribution for $k = 1, 2, \dots$ with mean k/β , if X is the sum of k independent random variables $X_1, X_2, \dots, X_k$ having a common exponential distribution with mean $1/\beta$. Particularly, it can be expressed as series of k exponential phases. Regarding this assumption, the packet latency contains k identical phases connected in series, each with exponentially distributed time. The mean duration of each phase is $\bar{X}/k$, where $\bar{X}$ denotes the mean of the whole time span. To divide the whole process of packet forwarding into k phases, the assumption is made that the series of packets that are moved from the `poll_list` to be processed one-by-one and to wait at Qdisc before they depart from the system have the relationship with the series of k phases. The assumption is reasonable because the time that a packet waits for processing plus processing time dominates the other waiting time. The study in Suksomboon et al. (2018) demonstrates that $k(s)$ depends on q given R, B, F_{cpu} are fixed. Then, it can be shorten as $k(q)$ and given by

$$k(q) = \lceil k(1)/q \rceil, \quad q = 2, 3, \dots, m \quad (6)$$



Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization, Fig. 3
Linux network stack-based packet forwarding for a single core server

where $k(1)$ is obtained by the curve fitting of the measurement result of configuring $q = 1$. Note that the formula in (6) is required in only one configuration.

Next, $1/\beta$ refers to the mean of an exponential distribution of variable X_i for all $i = 1, 2, \dots, k$. The factor that makes the mean of the variable in each phase scale due to changing q and R is the amount of time that the packets have to wait in the Rx DMA ring buffers. Thus, it can be assumed that β or $\beta(q)$ has a relationship with the total Rx hardware IRQ (Rx h-IRQ) rate, which is affected by these two parameters q and R . The received packets are uniformly distributed to all q ring buffers when the equally weighted flow director is applied. Thus, the packet arrival rate at each queue is equivalent to λ_{out}/q for weighted flow director $1/q$ since the packets were dropped before entering the queues. Define $\sigma(q, \lambda_{\text{in}})$ as the total Rx h-IRQ rate of all q Rx queues given λ_{in} and then the Rx h-IRQ rate of each Rx queue can be expressed as $\sigma(q, \lambda_{\text{in}})/q$. The Rx h-IRQ rate depends on the packet arrival rate at an Rx queue. For convenience, the Rx h-IRQ rate of each Rx queue is defined as $\psi(\lambda') = \sigma(q, \lambda_{\text{in}})/q$, where packet arrival rate at a Rx queue $\lambda' = \lambda_{\text{in}}/q$. Since the received packets are uniformly distributed to all q ring buffers and the Rx h-IRQ

rate of each queue depends on the packet arrival rate at each queue, $\psi(\lambda') = \sigma(q, \lambda_{\text{in}})/q$, where $\lambda_{\text{in}} = \lambda_{\text{out}}(q)/q$ and $\lambda_{\text{out}}(q)$ is the maximum throughput given q . Thus, with q Rx queues, the mean time of packets being waiting in their Rx DMA ring buffers can be rewritten as

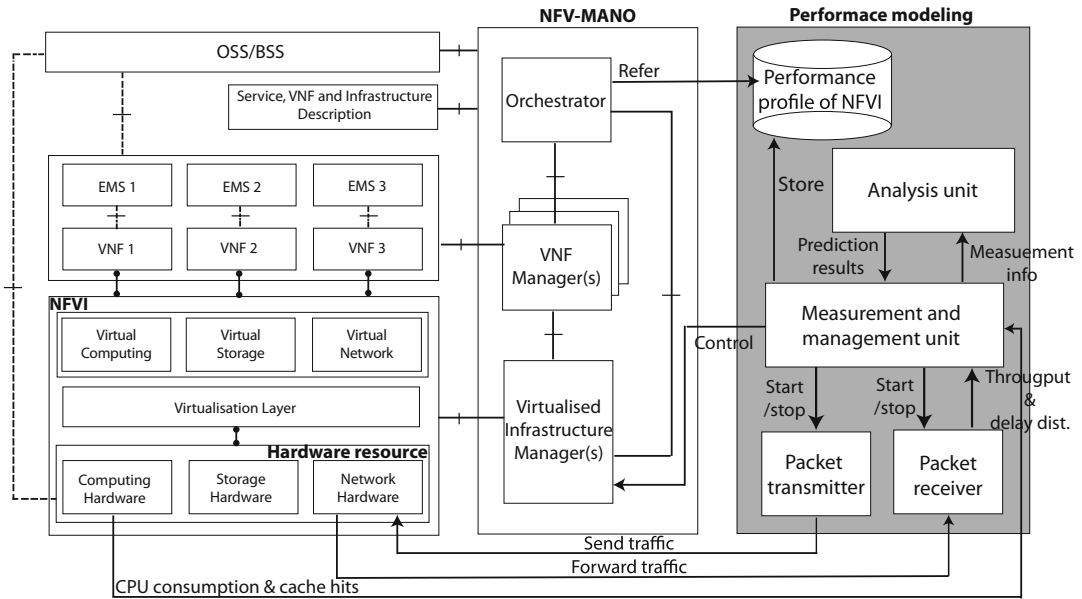
$$\frac{1}{\beta(q)} = q \left(\frac{\psi(\lambda_{\text{out}}(q)/q)}{\lambda_{\text{out}}(q)/q} \right) = \frac{q^2 \psi(\lambda_{\text{out}}(q)/q)}{\lambda_{\text{out}}(q)}. \quad (7)$$

Since $\gamma(q)$ implies the minimum time that a packet spending in a server, it can be intuitively estimated for each exponential phase by sampling an instance from the time of waiting for a Rx h-IRQ $\frac{q}{\psi(\lambda_{\text{out}}(q)/q)}$, and the time of the first packet arrives $\frac{1}{\lambda_{\text{out}}(q)/q}$. By assuming that these two terms are the independent random variables, their convolution is equal to their multiplication. This time spans according to the variance of the distribution, i.e., $k(q)/(\beta(q))^2$. That is $\gamma(q)$ is given by

$$\gamma(q) = \left(\frac{q^2 k(q)}{(\beta(q))^2 \psi(\lambda_{\text{out}}(q)/q) \lambda_{\text{out}}(q)} \right). \quad (8)$$

Key Applications

The abstraction of underneath resources with virtualization or network slicing opens new room



Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization, Fig. 4
Integration of performance modeling and NFV-MANO architecture

for mobile network businesses in which various sliced 5G services and virtual 5G appliances can be separately operated on top of these shared resources. This creation is expected to expand the telecommunications market to mobile virtual network operators (MVNOs) and let them enjoy a new playground of this sharing. It also helps in reducing the capital expenditure (CapEX) and operational expenditure (OpEX) of radio access network (RAN) and core network of mobile network operators (MNOs).

To allocate the virtual resources to each service slice optimally, the precise performance modeling of vPFF is a useful tool for assisting the NFV-MANO. Figure 4 presents the modules for the performance modeling of vPFF incorporated in NFV-MANO.

References

- Beifub A, Raumer D, Emmerich P, Raumer D, Runge T, Wohlfart F, Wolfinger B, Carle G (2015) A study of networking software induced latency. In: Proceedings of networked systems, pp 1–8
- Gallardo G, Baynat B, Begin T (2016) Performance modeling of virtual switching systems. In: Proceedings of IEEE MASCOTS, pp 125–134
- Gallenmüller S, Emmerich P, Wohlfart F, Raumer D, Carle G (2015) Comparison of frameworks for high-performance packet IO. In: Proceedings of IEEE ANCS, pp 29–38
- Kiasari A, Lu Z, Jantsch A (2013) An analytical latency model for networks-on-chip. *IEEE Trans VLSI Syst* 21:113–123
- Kim H, Kim K (2003) Performance analysis of the multiple input-queued packet switch with the restricted rule. *IEEE/ACM Trans Netw* 11:478–487
- Meyer T, Wohlfart F, Raumer D, Wolfinger B, Carle G (2014) Validated model-based performance prediction of multicore software routers. *IEEE Commun Surv Tutor* 37(1):93–107
- Mijumbi R, Serrat J, Gorricho J, Bouten N, DeTurck F, Boutaba R (2016) Network function virtualization: state-of-the-art and research challenges. *IEEE Commun Surv Tutor* 18(1): 236–262
- Riccobene V, McGrath M, Kourtis M, Xilouris G, Koumaras H (2016) Automated generation of VNF deployment rules using infrastructure affinity characterization. In: Proceedings of IEEE netsoft, pp 226–233
- Suksomboon K, Matsumoto N, Fukushima M, Shuichi O, Hayashi M (2017) Towards performance prediction of multicore software routers. *Int J Netw Manag* 27(2):1–19. <https://doi.org/10.1002/nem.1966>

Suksomboon K, Matsumoto N, Shuichi O, Hayashi M, Yusheng J (2018) Configuring a software router by the Erlang- k -based packet latency prediction. *IEEE J Sel Areas Commun* 36(3):1–16

Performance Prediction

- [Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization](#)

Permutation Modulation

- [Spatial Modulation](#)

Pervasive Computing Security

- [IoT Security](#)

Physical Layer Security

- [Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks](#)
- [Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices](#)

Pilot Contamination Attack

- [Pilot Spoofing Attack](#)

Pilot Reuse

- [Pilot Reuse for Massive MIMO](#)

Pilot Reuse for Massive MIMO

Li You and Xiqi Gao

National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Synonyms

[Channel acquisition](#); [Massive MIMO](#); [Pilot reuse](#); [Pilot scheduling](#)

Definitions

Pilot reuse refers to a pilot approach in wireless transmission where pilot signals adopted in the same cell can be non-orthogonal. Massive multiple-input multiple-output (MIMO) transmission employs a large number of antennas at the base station (BS) to serve a relatively smaller number of user terminals (UTs) simultaneously. For conventional orthogonal pilot approach, the pilot overhead scales linearly with the number of antennas, which constitutes the system bottleneck. Pilot reuse can be adopted to mitigate the pilot overhead issue in massive MIMO.

Background

Massive MIMO is a promising technology for future wireless systems due to its potential large gains in spectral efficiency and energy efficiency (Marzetta 2010; Wang et al. 2014). Channel state information (CSI) at the BS is of importance in massive MIMO transmission. In realistic systems CSI is typically obtained with the assistance of the periodically inserted pilot signals (Tong et al. 2004). For conventional orthogonal pilot approach, the pilot overhead is proportional to the number of antennas, which might become the system bottleneck for throughput in massive MIMO.

Massive MIMO channel properties can be adopted to address the above issue. Wireless channels are sparse in many typical propagation scenarios; most channel power is concentrated in a finite region of delays and/or angles due to limited scattering (Tse and Viswanath 2005; Liu et al. 2018; Xiao et al. 2017; Yang and Xiao 2018). Such channel sparsity can be resolved in the angle domain in massive MIMO due to the relatively large antenna array apertures (Payami and Tufvesson 2012; Gao et al. 2013). In addition, channel sparsity patterns, i.e., channel power distributions in the angle-delay domain, for different UTs are usually different. When pilots adopted by different UTs are properly scheduled according to the above channel properties, channel acquisition can be achieved simultaneously in an almost interference-free manner as with conventional orthogonal pilot approach (You et al. 2015, 2016; Zhong et al. 2015).

Massive MIMO Channel Model

Consider a single-cell wideband massive MIMO system, where the base station equipped with an M -element uniform linear array serves K single-antenna UTs. Orthogonal frequency division multiplexing (OFDM) modulation is adopted with the number of subcarriers being N_c and the sample length of the guard interval being N_g . Denote the uplink channel gain between the antenna of the k th UT and the antennas of the BS over OFDM symbol ℓ and subcarrier n as $\mathbf{g}_{k,\ell,n}$. Then the k th UT's channel at OFDM symbol ℓ over all subcarriers can be represented as $\mathbf{G}_{k,\ell} = [\mathbf{g}_{k,\ell,0} \ \mathbf{g}_{k,\ell,1} \ \dots \ \mathbf{g}_{k,\ell,N_c-1}]$. Motivated by the physical channel modeling approach (You et al. 2015, 2016), the angle-delay domain channel response matrix (ADCRM) of UT k at OFDM symbol ℓ can be defined as $\mathbf{H}_{k,\ell} = \mathbf{F}_M^H \mathbf{G}_{k,\ell} \mathbf{F}_{N_c \times N_g}^*$ where $\mathbf{F}_M$ denotes the $M \times M$ unitary discrete Fourier transform (DFT) matrix and $\mathbf{F}_{N_c \times N_g}$ denotes the matrix composed of the first N_g columns of the DFT matrix $\mathbf{F}_{N_c}$.

For OFDM-based massive MIMO channels, different elements of the ADCRM $\mathbf{H}_{k,\ell}$ are approximately mutually statistically uncorrelated (You et al. 2015, 2016), which lends the ADCRM its physical interpretation. Specifically, different elements of the ADCRM correspond to the channel gains for different incidence angles and delays, which can be resolved in massive MIMO-OFDM with a sufficiently large antenna array aperture. In addition, the statistics of the elements of ADCRM corresponds to the average channel power in the angle-delay domain and can be employed for pilot design in OFDM-based massive MIMO transmission.

Channel Acquisition with Pilot Reuse

Direct minimum mean square error (MMSE) estimation of the space-frequency domain channel response matrix $\mathbf{G}_{k,\ell}$ in massive MIMO is difficult to implement in practice due to the high complexity in large dimensional matrix inversion operation. However, channel estimation can be greatly simplified via properly exploiting the massive MIMO-OFDM channel properties. The BS can first estimate the ADCRM; then the space-frequency domain channel response matrix estimates can be readily obtained via exploiting the unitary equivalence between the angle-delay domain channels and the space-frequency domain channels while maintaining the same channel acquisition performance in terms of MSE.

Adjustable phase shift pilot (APSP) is one kind of pilot reuse approach for massive MIMO-OFDM channel acquisition (You et al. 2016). For the APSP approach, one sequence along with different adjustable phase-shifted versions of itself in the frequency domain is adopted as pilots for different UTs. APSPs are different from conventional phase shift orthogonal pilots (PSOPs) (Li 2002; Barhumi et al. 2003; Chi et al. 2011), in which phase shifts for different pilots are fixed and phase shift differences between different pilots are no less than the maximum channel delay (divided by the system sampling duration) of all the UTs. Since in the APSP

approach, the phase shifts for different pilots are adjustable, more pilots are available compared with conventional PSOPs, which leads to significantly reduced pilot overhead.

With APSPs, the sum MSE can be minimized when phase shifts for different pilots are properly scheduled. With frequency domain phase-shifted pilots, equivalent channels will exhibit corresponding cyclic shifts in the delay domain. If the equivalent channel power distributions in the angle-delay domain for different UTs can be made non-overlapping after pilot phase shift scheduling, the pilot interference effect can be eliminated, and the sum MSE of channel estimation can be minimized. However, such an optimal condition cannot always be met in practice, but pilot phase shift scheduling can still be beneficial. Pilot scheduling can be performed based on, e.g., MMSE criterion. The scheduling problem is combinatorial, and the optimal solution must be found through an exhaustive search. Motivated by the optimal condition for channel acquisition, simplified pilot phase shift scheduling algorithms can also be developed (You et al. 2015, 2016).

Key Applications

Pilot reuse can be adopted to mitigate the pilot overhead issue in massive MIMO. The idea of pilot reuse is motivated from the fact that, in realistic outdoor wireless propagation environments where the BS is located at a high altitude, the scattering around the BS is usually limited and the MIMO channels are not spatially isotropic. For UTs with channels lying in almost orthogonal spatial directions, pilot reuse is feasible and beneficial. For wideband massive MIMO-OFDM channels, APSP is one kind of pilot reuse approach that can efficiently reduce the pilot overhead and further improve the wireless transmission performance.

Cross-References

- [Wireless Fading Channel Model for 5G and Future](#)

References

- Barhumi I, Leus G, Moonen M (2003) Optimal training design for MIMO OFDM systems in mobile wireless channels. *IEEE Trans Signal Process* 51(6):1615–1624
- Chi Y, Gomaa A, Al-Dahir N, Calderbank AR (2011) Training signal design and tradeoffs for spectrally-efficient multi-user MIMO-OFDM systems. *IEEE Trans Wirel Commun* 10(7):2234–2245
- Gao X, Tufvesson F, Edfors O (2013) Massive MIMO channels – measurements and models. In: *Proceedings of the annual ASILOMAR*, Pacific Grove, pp 280–284
- Li YG (2002) Simplified channel estimation for OFDM systems with multiple transmit antennas. *IEEE Trans Wirel Commun* 1(1):67–75
- Liu Y, Fang X, Xiao M (2018) Discrete power control and transmission duration allocation for self-backhauling dense mmWave cellular networks. *IEEE Trans Commun* 66(1):432–447
- Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans Wirel Commun* 9(11):3590–3600
- Payami S, Tufvesson F (2012) Channel measurements and analysis for very large array systems at 2.6 GHz. In: *Proceedings of the EuCAP*, Prague, pp 433–437
- Tong L, Sadler BM, Dong M (2004) Pilot-assisted wireless transmissions: general model, design criteria, and signal processing. *IEEE Signal Process Mag* 21(6):12–25
- Tse D, Viswanath P (2005) *Fundamentals of wireless communication*. Cambridge University Press, New York
- Wang CX, Haider F, Gao XQ, You XH, Yang Y, Yuan D, Aggoune HM, Haas H, Fletcher S, Hepsaydir E (2014) Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Commun Mag* 52(2):122–130
- Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis G, Bjornson E, Yang K, Chih-lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. *IEEE J Sel Areas Commun (JSAC)* 35(9):1909–1935
- Yang G, Xiao M (2018) Performance analysis of millimeter-wave relaying: impacts of beamwidth and self-interference. *IEEE Trans Commun* 66(2):589–600
- You L, Gao XQ, Xia XG, Ma N, Peng Y (2015) Pilot reuse for massive MIMO transmission over spatially correlated Rayleigh fading channels. *IEEE Trans Wirel Commun* 14(6):3352–3366
- You L, Gao XQ, Swindlehurst AL, Zhong W (2016) Channel acquisition for massive MIMO-OFDM with adjustable phase shift pilots. *IEEE Trans Signal Process* 64(6):1461–1476
- Zhong W, You L, Lian T, Gao XQ (2015) Multi-cell massive MIMO transmission with coordinated pilot reuse. *Sci China Technol Sci* 58(12):2186–2194

Pilot Scheduling

► Pilot Reuse for Massive MIMO

Pilot Spoofing Attack

Wei Wang

ECE Department, University of Waterloo,
Waterloo, ON, Canada

Huazhong University of Science and
Technology, Wuhan, China

Synonyms

Pilot contamination attack; Spoofing attack

Definitions

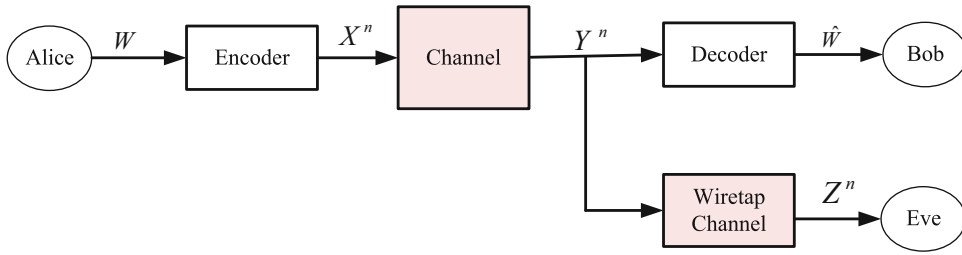
Pilot spoofing attack refers to an active attack in wireless networks, where active eavesdroppers (Eves) manipulate the channel estimation to increase the eavesdropping rate. In time division duplex (TDD) systems with channel reciprocity, the downlink channel state information (CSI) can be acquired through uplink channel training. A smart active Eve may hack such channel estimation process by sending an identical pilot sequence to the base station (BS), causing an erroneous channel estimation at the BS and more information leak to the Eve.

Background

By exploiting the randomness of wireless medium, physical layer security can provide lightweight security solution for low-cost Internet of Things (IoT) devices and has been regarded as a promising solution to enhance security in emerging large-scale heterogeneous networks (HetNets) (Mukherjee 2015; Poor and Schaefer 2017). The information-theoretical security was investigated by Wyner considering a wiretap

channel model in Wyner (1975), where Alice wants to transmit a confidential message W to Bob over a noisy channel while keep it secret to Eve, as shown in Fig. 1. Specifically, Alice encodes the message into a code word X^n with length n to enable Bob successfully decode W based on his observation Y^n . Wyner measured secrecy by equivocation or conditional entropy, and it is required that $\frac{1}{n}H(W|Z^n) = \frac{1}{n}H(W)$ to guarantee that W is kept secret from Eve. This criterion can also be equivalently written as $\lim_{n \rightarrow \infty} \frac{1}{n}I(W; Z^n) = 0$. Wyner characterized the secrecy capacity for the degraded wiretap channel and showed that secrecy capacity can be achieved through stochastic encoding. Since then, plenty of researches have been conducted. For the degraded wiretap channel with additive Gaussian noise, the secrecy capacity is determined by the capacity difference between the main channel and the wiretap channel Leung-Yan-Cheong and Hellman (1978). Hence, positive secrecy capacity is achievable only if the eavesdropping channel is a degraded version of the main channel. From the perspective of the Eve, to degrade the secrecy performance, advantages in the eavesdropping channel over the main channel can be realized by contaminating the channel estimation in the reverse channel training phase in TDD systems Zhou et al. (2012). Specifically, the Eve may send an identical pilot signal like a legitimate user in the uplink channel state information (CSI) acquisition phase and thus can flexibly control the eavesdropping rate by changing the uplink transmission power used for pilot spoofing attack. When the uplink transmission power of the Eve is large enough, channel estimation at the legitimate transmitter side will be dominated by the eavesdropping channel, and the probability of achieving positive secrecy rate will approach to zero.

Inspired by the pilot contamination phenomenon in multicell systems, pilot spoofing attack was firstly investigated in Zhou et al. (2012). It was shown that the effective signal-to-noise ratio (SNR) of the Eve may be higher than that of the legitimate receiver with proper power allocation for a full-duplex active Eve. To detect the pilot spoofing attack, an energy ratio



Pilot Spoofing Attack, Fig. 1 The Wyner wiretap channel model

detector was proposed in Xiong et al. (2015) by comparing the received signal power at the transmitter and receiver. By superimposing a random sequence onto the training sequence at the legitimate receiver and using the minimum description length (MDL) criterion, pilot contamination attack can be detected effectively (Tugnait 2015). More recently, a new random channel training (RCT) scheme was proposed in Wang et al. (2018a) to combat both the pilot spoofing and jamming attack. Through allocating multiple pilot sequences to a legitimate user and random pilot sequence selection, both the legitimate channel and eavesdropping channel can be estimated by exploiting the spatial channel statistics. However, this kind of scheme requires more training sequences and thus may not be applicable to multiple users case.

Due to the inherent similarity between pilot spoofing attack and pilot contamination in massive MIMO, the impact of pilot spoofing attack in massive MIMO has also been investigated. It was shown that the ergodic secrecy rate suffered a ceiling effect due to the pilot contamination in Zhu et al. (2014). In Basciftci et al. (2015), the authors revealed that the maximum secure degree of freedom (DoF) approaches to zero in the presence of active pilot spoofing attack. In Wu et al. (2016), by incorporating pilot contamination and pilot attack from multi-antenna active Eves in a TDD multicell multi-user massive MIMO system, a closed-form expression of the asymptotic achievable secrecy rate was derived with matched filter precoding and artificial noise. It was proved that the impact of the active Eve can be completely eliminated when the transmitting correlation matrices of the users and the

Eve are orthogonal, and robust design against pilot contamination attack was also developed. Furthermore, it was shown that deploying more base stations is more effective to improve the secrecy performance for a massive MIMO system under pilot contamination and pilot spoofing attack Wang et al. (2018b).

Impact of Pilot Spoofing Attack

To clarify the impact of pilot spoofing attack, we consider a single-cell multi-user massive MIMO system, where the BS is equipped with N antennas, whereas users and Eves are equipped with a single antenna each. We assume that each user is eavesdropped by an Eve. The channel from the k th user to the BS is denoted by $\sqrt{\beta_{u_k}}\mathbf{h}_k$, where β_{u_k} is the large-scale path loss and $\mathbf{h}_k$ is the small-scale Rayleigh fading. Similarly, the channel from the k th Eve is $\sqrt{\beta_{e_k}}\mathbf{g}_k$, with β_{e_k} being the large-scale path loss and $\mathbf{g}_k$ the small-scale fading vector.

We assume that the BS assigns orthogonal pilots for the K users within its cell, and Eve knows the training sequence assigned to the targeted user. Then the received signal at the BS in the uplink training stage can be expressed as

$$\mathbf{Y}_p = \sum_{k=1}^K \left(\sqrt{P_{u_k}\beta_{u_k}}\mathbf{h}_k\mathbf{s}_k^H + \sqrt{P_{e_k}\beta_{e_k}}\mathbf{g}_k\mathbf{s}_k^H \right) + \mathbf{N}, \quad (1)$$

where $\mathbf{s}_k$ is the pilot sequence assigned to the k th user with length τ_p and satisfies $\mathbf{s}_k^H\mathbf{s}_l = \delta_{kl}\tau_p$ with $\delta_{kl} = 1$ when $k = l$ and $\delta_{kl} = 0$ otherwise.

Note that P_{u_k} and P_{e_k} denote the uplink transmission power of the user and Eve, respectively, and $\mathbf{N}$ is the white Gaussian noise received at the BS. After correlating the received signal with the corresponding pilot sequence, the observed channel of the k th user can be expressed as

$$\mathbf{y}_k = \frac{\mathbf{Y}_p \mathbf{s}_k}{\tau_p} = \sqrt{P_{u_k} \beta_{u_k}} \mathbf{h}_k + \sqrt{P_{e_k} \beta_{e_k}} \mathbf{g}_k + \mathbf{n}_k. \quad (2)$$

Given the large-scale path loss, an MMSE channel estimation is given by Bai and Heath (2016) and Björnson et al. (2016)

$$\begin{aligned} \hat{\mathbf{h}}_k &= \frac{\sqrt{P_{u_k} \beta_{u_k}}}{\Sigma_k} \mathbf{y}_k = \frac{P_{u_k} \beta_{u_k}}{\Sigma_k} \mathbf{h}_k \\ &+ \frac{\sqrt{P_{u_k} P_{e_k} \beta_{u_k} \beta_{e_k}}}{\Sigma_k} \mathbf{g}_k + \frac{\sqrt{P_{u_k} \beta_{u_k}}}{\Sigma_k \Sigma_k} \mathbf{n}_k, \end{aligned} \quad (3)$$

where $\Sigma_k = P_{u_k} \beta_{u_k} + P_{e_k} \beta_{e_k} + \frac{\sigma^2}{\tau_p}$.

After uplink channel training, the BS uses the estimated channels to transmit downlink data to each user, and then the received signal at the k th user and k th Eve can be expressed as

$$\begin{aligned} y_k &= \sqrt{\phi_k P \beta_{u_k}} \mathbf{h}_k^T \mathbf{w}_k m_k \\ &+ \sum_{l \neq k}^K \sqrt{\phi_l P \beta_{u_k}} \mathbf{h}_k^T \mathbf{w}_l m_l + n_k, \end{aligned} \quad (4)$$

and

$$\begin{aligned} y_{e_k} &= \sqrt{\phi_k P \beta_{e_k}} \mathbf{g}_k^T \mathbf{w}_k m_k \\ &+ \sum_{l \neq k}^K \sqrt{\phi_l P \beta_{e_k}} \mathbf{g}_k^T \mathbf{w}_l m_l + n_{e_k}, \end{aligned} \quad (5)$$

respectively. Note that P is the total transmission power and ϕ_k , $\mathbf{w}_k$, and m_k are the power allocation, precoding vector, and intended message to the k th user, respectively. For a maximum-ratio transmission (MRT) scheme based on the estimated channel, i.e., $\mathbf{w}_k = \hat{\mathbf{h}}_k^* / \|\hat{\mathbf{h}}_k^*\|$, it can be seen that the desired signal for Eve satisfies $\lim_{N \rightarrow \infty} \frac{1}{N} |\mathbf{g}_k^T \mathbf{w}_k|^2 \rightarrow 0$ due to the correlation

between $\mathbf{g}_k$ and $\hat{\mathbf{h}}_k$. Therefore, the eavesdropping rate will not vanish by increasing the number of antennas in the presence of pilot spoofing attack.

Challenges and Open Problems

Considering the impact of pilot spoofing attack on the achievable secrecy performance, many challenges still need to be addressed:

- Pilot spoofing attack detection: Since the Eve uses the same pilot sequence with the legitimate user, the BS is usually unaware of the presence of pilot spoofing attack, which makes it more difficult to adopt countermeasure techniques. Therefore, effective detection schemes should be designed.
- Accurate legitimate channel estimation: To achieve secure transmission in the presence of pilot spoofing attack, advanced channel training schemes should be designed to obtain accurate channel estimations, especially in multiple Eves case. One possible approach is to exploit the uplink transmitted data and adopt the eigenvalue decomposition to estimate the channel eigenspace. Moreover, advanced double-training-based approaches may also be exploited to achieve accurate channel estimation.

Key Applications

Physical layer security is an appealing security solution for large-scale and resource-constrained networks, particularly in HetNet and machine-type communications, such as Internet of Things.

Cross-References

- [IoT Security](#)
- [Massive MIMO Channel Estimation](#)
- [Pilot Reuse for Massive MIMO](#)
- [Wireless Data Security](#)
- [Wireless Jamming Attack](#)

References

- Bai T, Heath RW (2016) Analyzing uplink SINR and rate in massive MIMO systems using stochastic geometry. *IEEE Trans Commun* 64(11):4592–4606
- Basciftci YO, Koksall CE, Ashikhmin A (2015) Physical layer security in massive MIMO. *arXiv preprint arXiv:150500396*
- Björnson E, Sanguinetti L, Kountouris M (2016) Deploying dense networks for maximal energy efficiency: small cells meet massive MIMO. *IEEE J Sel Areas Commun* 34(4):832–847
- Leung-Yan-Cheong SK, Hellman ME (1978) The Gaussian wire-tap channel. *IEEE Trans Inf Theory* 24(4):451–456
- Mukherjee A (2015) Physical-layer security in the internet of things: sensing and communication confidentiality under resource constraints. *Proc IEEE* 103(10):1747–1761
- Poor HV, Schaefer RF (2017) Wireless physical layer security. *Proc Natl Acad Sci* 114(1):19–26
- Tugnait JK (2015) Self-contamination for detection of pilot contamination attack in multiple antenna systems. *IEEE Wirel Commun Lett* 4(5):525–528
- Wang HM, Huang KW, Tsiftsis TA (2018a) Multiple antennas secure transmission under pilot spoofing and jamming attack. *IEEE J Sel Areas Commun* 36(4):860–876
- Wang W, Teh KC, Luo S, Li KH (2018b) Physical layer security in heterogeneous networks with pilot attack: a stochastic geometry approach. *IEEE Trans Commun*, to appear. <https://doi.org/10.1109/TCOMM.2018.2859954>
- Wu Y, Schober R, Ng DWK, Xiao C, Caire G (2016) Secure massive MIMO transmission with an active eavesdropper. *IEEE Trans Inf Theory* 62(7):3880–3900
- Wyner AD (1975) The wire-tap channel. *Bell Syst Tech J* 54(8):1355–1387
- Xiong Q, Liang YC, Li KH, Gong Y (2015) An energy-ratio-based approach for detecting pilot spoofing attack in multiple-antenna systems. *IEEE Trans Inf Forensic Secur* 10(5):932–940
- Zhou X, Maham B, Hjørungnes A (2012) Pilot contamination for active eavesdropping. *IEEE Trans Wirel Commun* 11(3):903–907
- Zhu J, Schober R, Bhargava VK (2014) Secure transmission in multicell massive MIMO systems. *IEEE Trans Wirel Commun* 13(9):4766–4781

Placement Strategies of Cellular Networks

- [Deployment Strategies of Cellular Networks](#)

Poisson Point Process

- [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)

Positioning-Locate-Locating

- [Location Based Service of Ad Hoc Networks](#)

Position-Locate

- [Location Based Service of Ad Hoc Networks](#)

Power Allocation

- [Power Control for Data Transmission in Wireless Networks](#)

Power Control

- [Power Control for Data Transmission in Wireless Networks](#)
- [Power Control in Wireless Networks](#)

Power Control for Data Transmission in Wireless Networks

Tuo Shi¹ and Zhipeng Cai²

¹School of Computer Science and Tech, Harbin Institute of Technology, Harbin, China

²Department of Computer Science, Georgia State University, Atlanta, GA, USA

Synonyms

Data transmission; Power allocation; Power control

Definitions

In wireless networks, data transmission should not be bothered by other interferences in the environment including background noise and interferences from other signals. Power control for data transmission is a primary resource allocation technique in wireless networks to guarantee successful data transmission. The power control technique has two goals. The first one is to provide each data signal proper quality without causing unnecessary interference to other signals. The second one is to reduce the total energy consumption in order to prolong network lifetime.

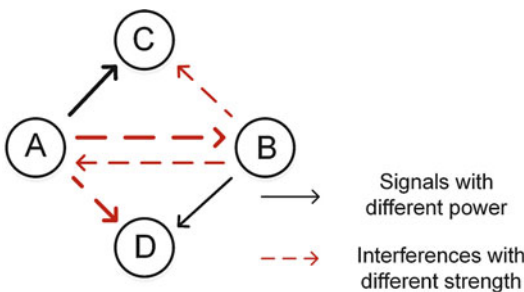
Historical Background

The interferences among different wireless transmission signals are major factors that influence the quality of data transmission in wireless networks. Once a node in a wireless network wants to transmit data to another node, the transmission signal will also cause interferences to its neighbors. As illustrated in Fig. 1, the transmission signals of nodes *A* and *B* interfere the data reception of other nodes. Node *A* may interfere nodes *B*, *C*, and *D*, and node *B* may interfere nodes *A*, *C*, and *D*. Furthermore, a signal with a high strength will cause more interferences. The ratio between signal and interference is named as Signal-to-

Interference Ratio (SIR) (Stanczak et al. 2009). In order to improve data transmission quality and make sure that data can be successfully transmitted, the SIR of each signal should satisfy some certain constraints.

In early years, i.e., 1980s or 1970s, some power control schemes were proposed aiming at balancing the SIR of data transmission links Aein (1973), Nettleton and Alavi (1983) and improving wireless link quality. The method in Nettleton and Alavi (1983) is only based on simple algebra, but it can increase link capacity by 30–100% compared with a non power control wireless system. The authors in Zander (1992) defined a measurement named *Interference Probability* to measure the performance of the power control algorithms. Suppose wireless data transmission requires an SIR threshold γ , under which data cannot be transmitted successfully, then the *Interference Probability* is the probability that the SIR is less than γ . As proved in Zander (1992), there is an upper bound of the *Interference Probability* achieved by all the power control algorithms, and the above SIR-balancing power control algorithms are optimal.

The above power control schemes all assume that the transmission power can be continuously adjusted and the energy capacity of each node is infinite. However, as mentioned in Grandhi and Zander (1994), such an assumption is very optimistic since the transmission nodes in a wireless network are always hand-held devices with limited energy capacity. Thus, in 1990s and 2000s, many authors Grandhi and Zander (1994), Shah et al. (1998), Goodman and Mandayam (2000), Xiao et al. (2001) investigated how to improve the SIR of signals and optimize energy consumption as well. The work in Grandhi and Zander (1994) considers a maximum transmitted power P_{max} , and the SIR is minimized with the power of each signal being less than P_{max} . Other works, such as Shah et al. (1998), Goodman and Mandayam (2000), Xiao et al. (2001, 2003), Saraydar et al. (2002), consider network utility functions to improve data transmission quality and reduce energy consumption. Although the utility function proposed in each work is different, all the works use SIR as a major metric and energy



Power Control for Data Transmission in Wireless Networks, Fig. 1 Interferences between different nodes. The transmission signals of nodes *A* and *B* interfere the data reception of other nodes. Node *A* may interfere nodes *B*, *C*, and *D*, and node *B* may interfere nodes *A*, *C*, and *D*

consumption as a penalty. The power control process is formulated as a noncooperative game in these papers, and each node in a wireless network is treated as a player and tries to maximize its own utility. Thus, if the game researches a Nash equilibrium, then the global utility of the network is maximized.

The above methods all aim at improving signal quality. However, the authors in Yates (1995) and Li and Ephremides (2007) recommended that instead of optimizing signal capacity, prolonging node lifetime is more important. Thus, in Yates (1995) and Li and Ephremides (2007), some power control methods have been proposed to minimize the total energy consumption and guarantee a proper quality of each signal.

Although power control for wireless data transmission has been investigated in several decades, it is still an important research area in the recent years. The authors in Karipidis et al. (2015) proposed an SIR-balancing method with the assumption that the receivers in a network have an interference cancelation capability. It means the receivers can decode interfering signals whenever these interfering signals are strong enough. Zappone et al. (2016) and (2018) proposed some power control methods in 5G wireless networks.

Key Applications

Power control for wireless data transmission is involved in the areas related to wireless communications, such as communications among satellites, mobile cellular network users, and so on. In the future, it can also be used in 5G wireless networks, edge computing techniques, and so on.

Cross-References

- [5G Wireless](#)
- [Energy Efficiency of Cellular Networks](#)
- [Interference-Aware Distributed Resource Allocation](#)

- [Multiple Access Technique for Cellular Wireless Networks](#)

References

- Aein J (1973) Power balancing in systems employing frequency reuse. *COMSAT Tech Rev* 3:277
- Goodman D, Mandayam N (2000) Power control for wireless data. *IEEE Pers Commun* 7(2):48–54
- Grandhi SA, Zander J (1994) Constrained power control in cellular radio systems. In: Vehicular technology conference, 1994 IEEE 44th, IEEE, pp 824–828
- Karipidis E, Yuan D, He Q, Larsson EG (2015) Max-min power control in wireless networks with successive interference cancelation. *IEEE Trans Wirel Commun* 14(11):6269–6282
- Li Y, Ephremides A (2007) A joint scheduling, power control, and routing algorithm for ad hoc wireless networks. *Ad Hoc Netw* 5(7):959–973
- Nettleton R, Alavi H (1983) Power control for a spread spectrum cellular mobile radio system. In: Vehicular technology conference, 1983, 33rd IEEE, IEEE, vol 33, pp 242–246
- Saraydar CU, Mandayam NB, Goodman DJ (2002) Efficient power control via pricing in wireless data networks. *IEEE trans Commun* 50(2):291–303
- Shah V, Mandayam NB, Goodman DJ (1998) Power control for wireless data based on utility and pricing. In: Personal, indoor and mobile radio communications, 1998. The Ninth IEEE international symposium on, IEEE, vol 3, pp 1427–1432
- Stanczak S, Wiczanowski M, Boche H (2009) Fundamentals of resource allocation in wireless networks: theory and algorithms, vol 3. Springer Science & Business Media, Berlin
- Xiao M, Shroff NB, Chong EK (2001) Utility-based power control in cellular wireless systems. In: INFOCOM 2001. Twentieth annual joint conference of the IEEE computer and communications societies. Proceedings. IEEE, IEEE, vol 1, pp 412–421
- Xiao M, Shroff NB, Chong EK (2003) A utility-based power-control scheme in wireless cellular systems. *IEEE/ACM Trans Networking (TON)* 11(2):210–221
- Yates RD (1995) A framework for uplink power control in cellular radio systems. *IEEE J Sel Areas Commun* 13(7):1341–1347
- Zander J (1992) Performance of optimum transmitter power control in cellular radio systems. *IEEE Trans Veh Technol* 41(1):57–62
- Zappone A, Sanguinetti L, Bacci G, Jorswieck E, Debbah M (2016) Energy-efficient power control: a look at 5G wireless technologies. *IEEE Trans Signal Process* 64(7):1668–1683
- Zappone A, Sanguinetti L, Debbah M (2018) Energy-delay efficient power control in wireless networks. *IEEE Trans Commun* 66(1):418–431

Power Control in Wireless Networks

Lei Yang

University of Nevada Reno, Reno, NV, USA

Synonyms

Power Control; Wireless Networks

Definitions

Power control in wireless networks, broadly speaking, is the intelligent control of transmitters' power output to achieve good system performance, such as network capacity, network connectivity, and lifetime of network devices and even the network.

Historical Background

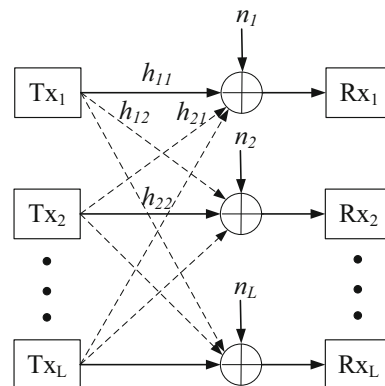
Power control in wireless networks has been systematically studied since the 1970s. Over the last two decades, wireless networks (e.g., cellular networks and WiFi networks) have been exponentially growing and making transformative impacts on society. Due to the broadcast nature of wireless communication, wireless networks are susceptible to interference, particularly in interference-limited systems, such as CDMA systems. Power control helps to mitigate mutual interference among the users, so as to ensure efficient spectral reuse and desirable user experience. Besides interference management, power control can help to minimize the overall energy consumption, which is particularly important for the lifetime of mobile devices with limited battery power. Moreover, due to the uncertainty and time-varying nature of wireless channels, power control is important for maintaining the connectivity between the transmitter and the receiver by ensuring a minimum level of received signal. Extensive research on wireless network power control has produced a wide and deep set of

results in terms of modeling, analysis, and design (see Chiang et al. 2008; Gunnarsson 2004).

Foundations

Power control in wireless networks adopts a radio link model, in which each link consists of a dedicated transmitter-receiver pair and the transmitted power of each link appears as interference to other links (see Fig. 1). Let p_l be the transmit power of link l and h_{kl} be the channel gain between link k 's transmitter and link l 's receiver. The receiver of link l receives the signal at the power of $h_{ll}p_l$ and is subject to the total interference $\sum_{k \neq l} h_{kl}p_k$ and the noise n_l . The quality of the received signal of link l is quantified by the signal-to-interference-plus-noise ratio (SINR), i.e., $\gamma_l = h_{ll}p_l / (\sum_{k \neq l} h_{kl}p_k + n_l)$.

A power control problem often involves an optimization formulation with transmit power as optimization variables and aims to optimize system utility or resource cost subject to various constraints (e.g., total transmit power, maximum transmit power for each link). For example, an increasing convex function (e.g., linear function) of transmit power can be used as the objective function. For system utility, QoS-based utility function is often used and assumed to be additive across links, i.e., $U(\boldsymbol{\beta}) = \sum_l U_l(\beta_l)$, where $\boldsymbol{\beta}$ is a vector of metrics. One commonly used QoS



Power Control in Wireless Networks, Fig. 1 Radio link model, where “Tx” and “Rx” represents transmitter and receiver, respectively (Yang et al. 2012)

metric is throughput, which is a function of SINR and in turn a function of transmit powers. The throughput is often assumed to be the Shannon rate achievable over Gaussian flat, fading channels, i.e., $\beta_l = w \log(1 + c\gamma_l)$, where w is the bandwidth in Hertz and c is a constant depending on modulation scheme, symbol period, and target bit error rate (BER).

Different formulations for the power control problem have been studied, which can be generally categorized into two types: optimization and game theory approach. In the game theory approach, all links compete for limited radio resources, and the problem is formulated and solved as a noncooperative game in MacKenzie and Wicker (2001), where the solution is a Nash equilibrium power allocation, at which no link has incentive to unilaterally change its power. In contrast, the optimization approach assumes that all links update their powers for social welfare maximization in Lee et al. (2006), Tse and Viswanath (2005), and Chiang et al. (2007).

Efficiently solving the power control problem depends on whether the problem is convex or not. As convexity is long recognized as the watershed between easy and hard optimization problems (see Rockafellar 1993), the power control problem is often formulated as convex optimization, i.e., minimization of a convex objective function over a convex constraint set (see Lee et al. 2006; Tse and Viswanath 2005; Chiang et al. 2007; Foschini and Miljanic 1993; Hande et al. 2008). When the problem is formulated as nonconvex optimization (e.g., a discrete set of available power levels in Wu and Bertsekas 2001 and nonconvex objective functions in Yang et al. 2012; Qian et al. 2009), the problem becomes highly difficult to solve. For certain nonconvex cases, the problem can be still solved efficiently. For example, a log change of variables can turn an apparently nonconvex problem into a convex one as in the geometric programming approach that has been shown to solve a wide range of constrained power control problems in high SINR regime and a smaller set of problems in general SINR regime (see Chiang et al. 2007).

While convexity is the key to global optimality and efficient computation, decomposabil-

ity is the key to distributed solutions of the power control optimization problem. It is desirable to decompose the power control problem into subproblems with no communication overhead or message passing. In noncooperative game formulations of power control, distributed algorithms are naturally provided, such that each link can update its power by monitoring the total interference plus noise. In the optimization approach, various decomposition techniques, such as dual decomposition, primal decomposition, and penalty function method, can be used to decompose the original problem.

Key Applications

Fixed SINR The basic problem formulation for cellular network power control has been extensively studied, where transmit powers are the only variables, constrained by fixed target SINR γ_l and optimized to minimize the total power. To compute an optimal solution, the distributed power control (DPC) algorithm is proposed by Foschini and Miljanic in (1993), where the transmit power of link l is updated by $p_l(t+1) = \gamma_l p_l(t) / \text{SINR}_l(t)$, in which $t = 1, 2, \dots$, and $\text{SINR}_l(t)$ is the received SINR at the t th iteration. This algorithm is distributed as each link only needs to monitor its individual received SINR and can update its transmit power independently and asynchronously (see Yates 1995). It has been shown that this algorithm can achieve the optimal power allocation in the limit as t goes to infinity.

Variable SINR The fixed SINR approach is suitable for lightly loaded voice networks. When the network is heavily loaded, a wireless network needs to optimize the SINR assignment according to traffic requirements and channel conditions. Therefore, the problem becomes a joint SINR assignment and power control. In the vast research literature since the mid-1990s, solutions to this joint optimization problem are either distributed but suboptimal or optimal but centralized. For convex formulations of this problem, the work in Hande et al. (2008) develops

a distributed and optimal algorithm. The main idea is to reparametrize the complicated, and coupled constraint set into the so-called load-spillage representation, essentially a sophisticated change of coordinates, based on which the load-spillage power control (LSPC) algorithm has been developed such that each mobile user can compute an ascent search direction for this optimization problem locally. Convergence and optimality of this algorithm have been provided. Recently, LSPC has been adopted in the industry.

Future Directions

With the rapid increase of data service, the demand for higher data rate in wireless networks has been continuously increasing. Many wireless communication technologies have been developed, e.g., full duplex (FD) communication and device-to-device (D2D) communication, aiming to provide higher speed, larger capacity, and better QoS. At the same time, new challenges have been introduced. For example, FD communication will introduce strong self-interference, and D2D communication will interfere with not only the cellular communication but also the communication between other mobile devices. In the future, power control in wireless networks will focus on addressing these challenges, and it is expected that new modeling techniques, theoretical advances, and mathematical machinery will be required for these wireless and mobile applications.

Cross-References

- [Interference-Aware Distributed Resource Allocation](#)
- [Interference-Aware Scheduling in Wireless Networks](#)
- [Interference Characterization for Cooperative Communications in Mobile Ad Hoc Networks](#)

References

- Chiang M, Tan CW, Palomar DP, O'Neill D, Julian D (2007) Power control by geometric programming. *IEEE Trans Wirel Commun* 6(7):2640–2651
- Chiang M, Hande P, Lan T, Tan CW, et al (2008) Power control in wireless cellular networks. *Found Trends@ Netw* 2(4):381–533
- Foschini GJ, Miljanic Z (1993) A simple distributed autonomous power control algorithm and its convergence. *IEEE Trans Veh Technol* 42(4): 641–646
- Gunnarsson F (2004) Power control in wireless networks. In: Guizani M (ed) *Wireless communications systems and networks*. Springer, Boston, pp 179–208
- Hande P, Rangan S, Chiang M, Wu X (2008) Distributed uplink power control for optimal sir assignment in cellular data networks. *IEEE/ACM Trans Netw (TON)* 16(6):1420–1433
- Lee JW, Mazumdar RR, Shroff NB (2006) Opportunistic power scheduling for dynamic multi-server wireless systems. *IEEE Trans Wirel Commun* 5(6):1506–1515
- MacKenzie AB, Wicker SB (2001) Game theory in communications: motivation, explanation, and application to power control. In: *Global telecommunications conference, 2001, GLOBECOM'01*, vol 2. IEEE, Piscataway, pp 821–826
- Qian LP, Zhang YJ, Huang J (2009) Mapel: achieving global optimality for a non-convex wireless power control problem. *IEEE Trans Wirel Commun* 8(3):1553–1563
- Rockafellar RT (1993) Lagrange multipliers and optimality. *SIAM Rev* 35(2):183–238
- Tse D, Viswanath P (2005) *Fundamentals of wireless communication*. Cambridge University Press, Cambridge
- Wu C, Bertsekas DP (2001) Distributed power control algorithms for wireless networks. *IEEE Trans Veh Technol* 50(2):504–514
- Yang L, Sagduyu YE, Zhang J, Li JH (2012) Distributed stochastic power control in ad hoc networks: a non-convex optimization case. *EURASIP J Wirel Commun Netw* 2012(1):231
- Yates RD (1995) A framework for uplink power control in cellular radio systems. *IEEE J Sel Areas Commun* 13(7):1341–1347

Power Delay Profile

- [Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System](#)

Power Domain NOMA

- [Non-orthogonal Multiple Access](#)

Power Harvesting Technology

- [Energy Harvesting Technologies in Wireless Sensor Networks](#)

Power Line Communications

- [Interference Alignment for Power Line Communications](#)

Power Line Communications for Grid Discovery and Diagnostics

Federico Passerini and Andrea M. Tonello
Institute of Networked and Embedded Systems,
Alpen-Adria-Universität Klagenfurt, Klagenfurt,
Austria

Synonyms

[Grid monitoring and anomaly detection using power line communications](#); [Grid surveillance](#)

Definitions

Grid discovery and diagnostics refers to characterizing both the electrical and topological properties of the grid and to monitor their evolution over time. It can be performed using two technical approaches. The first approach employs a series of electrical meters that are directly branched to the network and specifically designed to sense different electrical quantities.

The gathered information is then shared between the nodes or directly with a central unit using a separate communication device. The second approach exploits power line communications both as sensing and communication technology. Power line communications (PLC) transmitters generate signals according to their specific communication standard, while the receivers, apart from the classical communication function, exploit those signals to perform grid discovery and diagnostics. The sensing signals have a frequency range from few Hz to few kHz in the first approach, while in the second, as per PLC standards, the frequency range goes from few kHz to several MHz.

Historical Background

Grid discovery and diagnostics are traditionally performed using two separate devices. The first is a dedicated hardware, normally an electrical meter, that is branched to the grid and continuously senses the network. The second is a communication device endowed with a modem that is used to share the sensed data and possible control commands. The most common measurement device is called phasor measurement unit (PMU). Many PMUs are normally deployed across the grid and measure in a synchronized fashion the values of the main voltage and current about 30 times per second. When required, they can reach up to few thousand measurements per second. Concerning communications, there are different possibilities, including PLC, radio system, fiber optic, and satellite communications (Chairman et al. 2011). This monitoring system is well established in the context of high-voltage transmission networks and in wide medium-voltage distribution networks. However, conventional PMUs are not usually deployed in small-sized distribution networks or in micro-grids. In fact, given the average length of the power cables in such networks, frequencies higher than few kHz are preferable to efficiently perform grid discovery and diagnostics. In such networks utilities have a very low level of control.

For these reasons, a novel stream of research is proposing to use the high frequency of PLC to absolve discovery and diagnostic tasks in small-size grids. This has a double advantage. On one hand, with high frequency measurements, greater resolution can be obtained. On the other hand, since PLC are both used to monitor the network and to send communication signals, they embed in a single technology the role of sensors and modems. The idea is similar to what has been done for Internet access networks. In that context, digital subscriber line (DSL) modems have been used since the 1990s to monitor the “last mile” of telephone networks. The characteristics of power networks are, however, completely different, which requires a separate treatment.

Background on Grid Discovery

Grid discovery has emerged as a practical and research problem with the advent of convoluted and interconnected networks. The term refers to the ability of an ensemble of network nodes to infer the physical structure and the electrical parameters of the network. The nodes are enabled to discover their relative position within the grid, to detect the presence of other nodes, and, ultimately, to derive the grid topology. Furthermore, they are enabled to infer the electrical parameters of the cables, the impedance value of the loads, and the current and voltage at each termination. A complete knowledge of the structure of the grid allows utilities to optimize different aspects: the flow of the delivered power, the values of the balancing loads, the routing strategies of information signals for control, and the defensive strategies against possible intruders.

Grid discovery is normally treated as a state estimation problem based on the data provided by PMUs. The produced measurements are used in combination with state estimation algorithms to infer, depending on the application, different properties of the network (Huang et al. 2012). In transmission networks, it is of interest to estimate the voltage and current phasors at every node. In micro-grids, on the other hand, it is of interest to monitor the topological changes of the grid, since they occur too often to be manually controlled by an operator. Such techniques are always based on

complete or at least partial knowledge of the set of the possible topologies of the grid.

In recent years, interest has grown toward the inference of the topology of a distribution or micro-grid with no previous information about it. Some approaches, presented in the following, exploit high-frequency signals provided by PLC modems. Their use is of particular interest in medium- and low-voltage distribution networks, where the utilities have limited knowledge of the actual structure of the network.

Background on Grid Diagnostics

Grid diagnostics involve the constant monitoring of the grid with electrical signals in order to detect and possibly localize possible anomalies that might jeopardize the distribution of electrical power. Every consequent damage on the grid has to be repaired within the shortest time possible, in order to maintain a high service quality for the costumers. However, since transmission and distribution grids cover wide areas and they are often rather convoluted, detecting and localizing anomalies in a short time is a complex task. Given the huge economical and social impact that strong anomalies like short circuits can have, grid diagnostics have been performed on power grids since the beginning of the twentieth century. The proposed monitoring techniques have since then constantly evolved.

Grid diagnostics are typically based on power flow analysis (Grainger et al. 2016), and it focuses on the detection and localization of electrical faults. The basic technique consists of deploying one or more relays across the grid to monitor the delivered voltage and the total current absorbed by the loads. The presence of a fault is detected when the expected values of voltage and current exceed a predetermined threshold. The location of the fault is determined analyzing the phase of the fault current. Older systems exploited just mechanical relays that used fuses to detect overcurrents and trigger a circuit breaker. Modern microprocessor-enabled evolutions of these systems are still applied nowadays, especially in transmission networks (Das et al. 2014).

Along the years, more complex analysis techniques have been layered on basic threshold methods in order to detect and localize faults in topologically complex networks, to classify different kind of events, and to detect them even if they are very short or very weak (Shafiullah and Abido 2017; Sedighizadeh et al. 2010). Such techniques make use of various tools that can be broadly divided into deterministic and heuristic approaches. Deterministic approaches exploit time-domain or frequency-domain analysis applied to each new sensing event. Heuristic approaches, like neural or fuzzy networks and genetic algorithms, are based on training data and the available information about the grid. They are able to automatically detect and localize faults based on all the system history. Conversely from deterministic approaches, they require much more computation power, but they can better adapt to system changes.

In recent years, a lot of interest has grown toward diagnostics in underground grids and small-scale overhead grids which, as previously discussed, are not well controlled by the utilities. Given the small dimension of these grids, monitoring at high frequency is needed. For this reason, a series of methods that make use of PLC modems has been proposed not only to detect faults but also to track the aging of cables (particularly important in underground grids) and eventual changes of the loads. In the following, a description of these methods is presented.

Foundations

In order to optimize the communication rate, every communication protocol requires to perform channel estimation in order to apply equalization algorithms at the receiver. The channel is composed by the physical medium, the couplers, and the analog part of the transceivers. In the context of PLC, the medium is the power grid. If the couplers and the analog transceiver at the transmitter and at the receiver are well characterized, it is possible to estimate the medium transfer function (MTF). The MTF does not depend on the transmitted signals (although its estimation

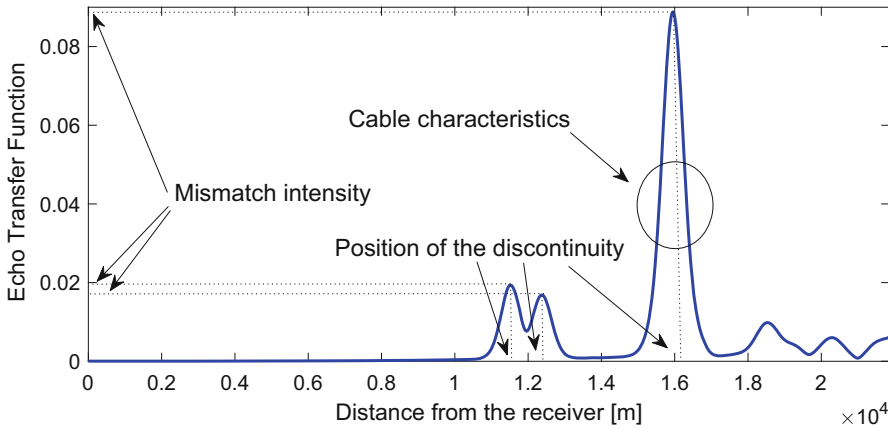
depends on it), but it is an intrinsic property of the medium.

Each receiving modem can estimate the link transfer function (LTF) with respect to a transmitting modem. When the receiving modem starts a communication with another transmitter, then a new LTF is estimated, referring to the new link. If a modem is equipped with an impedance measurement circuit, it can also estimate the echo transfer function (ETF), i.e., from the transmitter modem to itself. Conversely from the LTF, the ETF is an intrinsic property of the node the modem is branched to.

The LTF and the ETF contain all the structural and electrical information about a specific power grid. The analysis of such information is based on the transmission line theory (Paul 2007; Passerini and Tonello 2018b). For example, from the time-domain ETF depicted in Fig. 1, it is possible to infer the following information. Each peak determines the position of a discontinuity in the grid, which might be a cable branch, a termination node, or even an anomaly. The amplitude of each peak is determined by the amount of the mismatch of the discontinuity with the line. This ultimately depends on the characteristic impedance of the power line and on the equivalent impedance of the load or the anomaly. Finally, the shape and the amplitude of the peaks depend on the electrical characteristics of the power line. Due to signal attenuation, frequency, and time spread, the access to this information is often limited, especially regarding the parts of the grid that are distant from the measurement point. For this reason, it is often preferred to fuse the information provided by the LTF of multiple links or the ETF of multiple nodes. Furthermore, the closer the nodes are to each other, the higher is the frequency bandwidth required to distinguish them.

Application: Grid Discovery

Three main approaches have been proposed to date to perform grid discovery, i.e., to estimate the grid topology. The first approach is based on reflectometry, that is, the estimation of the



Power Line Communications for Grid Discovery and Diagnostics, Fig. 1 Example of a time-domain ETF and associated information about the medium

echo signal from one modem deployed on a network termination. Assuming the presence of no anomaly, the peaks in the time-domain ETF unveil the presence of either branching points or terminations. Different algorithms have been proposed to emulate the peak sequence of the ETF by guessing node by node (Zhang et al. 2016) or peak by peak (Ahmed and Lampe 2012) the topology of the grid. This approach has the advantage of relying on just one measurement point, but the grid discovery is limited to few nodes next to it.

The second approach avoids peak analysis by operating on the data-link layer. A set of modems is deployed on all the grid terminations (Lampe and Ahmed 2013), or even on every node (Erseghe et al. 2013). Every modem has to establish a connection with each one of the others. During the handshake phase, the modem is capable of computing the round-trip time and, thus, to estimate the length of the link. When the length of all the possible links in the network is known, different algorithms based on decision and graph theories can be applied to infer the topology. This approach has the advantage of being scalable to high complex networks but requires an extensive deployment of broadband PLC modems.

The third approach relies on impedance measurements performed on the network nodes (Passerini and Tonello 2017). If the user has some

prior information about the cable parameters and the loads, it is possible to infer both the node-to-node distances and the topology in a unified algorithm by using the classical propagation equations from transmission line theory. This approach requires a lot of prior information and extensive deployment of PLC modems but can work both with wide-band measurements or using a single frequency down to few tenths of kHz.

A summary of the grid discovery methods using PLC is shown in Table 1.

Application: Grid Diagnostics

High-frequency grid diagnostics can be performed when at least one reference measurement, performed during the normal operation of the grid, is available. Then, the medium is continuously monitored, and every new estimation of the MTF is subtracted from or divided by the reference measurement, depending on the system model used (Passerini and Tonello 2018b), providing a test trace. Since the power line medium is periodically time variant in the 3 kHz–10 MHz range, using just one reference measurement would often provide non-null test traces. In order to discover anomalies, it has been proposed to analyze the features of the test traces with a machine learning algorithm. Such

Power Line Communications for Grid Discovery and Diagnostics, Table 1
Grid discovery methods using PLC

Method	Class	Advantages	Disadvantages
Reflectometry and model-based search	Echo	PLC modem on just one node, few prior information needed	Topology inference limited to few branches, requires wide band
Handshake and inference algorithm	Link	Scalable, few prior information needed	PLC modem on every termination or even on every node, requires wide band
Impedance	Echo	Single or multiple frequencies allowed, works also using low frequencies, cable length and connections in a single algorithm	Many parameters required, PLC modem possibly on every node

Power Line Communications for Grid Discovery and Diagnostics, Table 2
Grid diagnostics methods using PLC

Method	Class	Advantages	Disadvantages
Deterministic LTF	Link	Detection both close to the transmitter and the receiver	Ambiguity on the location
Heuristic LTF	Link	Detection both close to the transmitter and the receiver	Requires huge database, ambiguity on the location
Deterministic ETF	Echo	Reduced complexity algorithms, univocal localization	Detection only close to the sensing modem

algorithm has been tested to track the aging of underground cables, providing promising results (Förstel and Lampe 2017).

On the other hand, considering the periodicity of the channel variation as composed by a series of invariant states, it is possible to use an equal number of reference measurements. This way, if no anomaly is present, the amplitude and phase of the resulting trace is limited to a noise component. Every time a threshold set consistently to the average noise power level is exceeded, an anomaly is detected. This approach has been presented considering a single LTF or a single ETF (Passerini and Tonello 2018a), but of course the detection performance is improved when relying on measurements from multiple node or links. Different classification algorithms based on the features of the test trace can then be conceived to identify the type of the detected anomaly (Passerini and Tonello 2018a).

For what concerns the localization of an anomaly, it is performed by analyzing the test trace in time domain, and it mostly relies on

the knowledge of the network topology. When the topology of the grid is not known, it is only possible to retrieve the distance of the anomaly from a communication node, which corresponds to the position of the first peak in the time-domain MTF (either LTF or ETF). In the case of the ETF, the distance is univocally referred to the measurement point, but in the case of the LTF, an ambiguity is left, since the distance might be either referred to the transmitter node or to the receiver node (Passerini and Tonello 2018b). When the topology of the grid is known, then the precise branch where the anomaly takes place can be identified by a peak analysis of the test trace.

A summary of the grid diagnostic methods using PLC is shown in Table 2.

Cross-References

- [Advances in Distribution System Monitoring](#)
- [Scalable Smart Grid Communication](#)

References

- Ahmed M, Lampe L (2012) Power line network topology inference using frequency domain reflectometry. In: 2012 IEEE International Conference on Communications (ICC), pp 3419–3423
- Chairman et al (2011) Digital communications for relay protection. Working Group H9 of the IEEE Power System Relaying Committee. [http://www.pes-psrc.org/kb/published/reports/Digital%20communications%20for%20relaying%20\(H9\).pdf](http://www.pes-psrc.org/kb/published/reports/Digital%20communications%20for%20relaying%20(H9).pdf)
- Das S, Santoso S, Gaikwad A, Patel M (2014) Impedance-based fault location in transmission networks: theory and application. *IEEE Access* 2:537–557
- Erseghe T, Tomasin S, Vigato A (2013) Topology estimation for smart micro grids via powerline communications. *IEEE Trans Signal Process* 61(13):3368–3377
- Förstel L, Lampe L (2017) Grid diagnostics: monitoring cable aging using power line transmission. In: 2017 IEEE International Symposium on Power Line Communications and its Applications (ISPLC), pp 1–6
- Grainger J, Stevenson W, Chang G (2016) Power system analysis. McGraw-Hill Education, New York
- Huang YF, Werner S, Huang J, Kashyap N, Gupta V (2012) State estimation in electric power grids: meeting new challenges presented by the requirements of the future grid. *IEEE Signal Process Mag* 29(5): 33–43
- Lampe L, Ahmed M (2013) Power grid topology inference using power line communications. In: 2013 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp 336–341
- Passerini F, Tonello AM (2017) On the exploitation of admittance measurements for wired network topology derivation. *IEEE Trans Instrum Meas* 66(3): 374–382
- Passerini F, Tonello AM (2018a) Smart grid network sensing using power line modems: anomaly detection and location. Available on arXiv. <https://arxiv.org/abs/1807.05347>
- Passerini F, Tonello AM (2018b) Smart grid network sensing using power line modems: effect of anomalies on signal propagation. Available on arXiv. <https://arxiv.org/abs/1806.10991>
- Paul CR (2007) Analysis of multiconductor transmission lines. Wiley-IEEE Press, New York
- Sedighzadeh M, Rezazadeh A, Elkalashy I (2010) Approaches in high impedance fault detection – a chronological review. *Adv Electr Comput Eng* 10(3):114–128
- Shafiullah M, Abido MA (2017) A review on distribution grid fault location techniques. *Electr Power Compon Syst* 45(8):807–824
- Zhang C, Zhu X, Huang Y, Liu G (2016) High-resolution and low-complexity dynamic topology estimation for PLC networks assisted by impulsive noise source detection. *IET Commun* 10(4):443–451

Preferences

► Preferences and Utility Functions

Preferences and Utility Functions

Randall A. Berry
Department of Electrical Engineering and
Computer Science, Northwestern University,
Evanston, IL, USA

Synonyms

[Preferences](#); [Valuations](#); [Willingness-to-pay](#)

Definition

Utility functions are a mathematical representation of an individual's preferences for different goods and services.

Motivation

The notion of utility is motivated by a basic problem faced when allocating limited resources (such as bandwidth or power) to users of a wireless network: How should one account for differences in the users' needs for these resources? These differences can arise due to users running different applications that require different quality of service (QoS) levels or even for the same application due to differences in how users perceive changes in QoS. Moreover, modern wireless networks exist in evolving market structures with multiple competing firms using heterogeneous technologies. Understanding how this competition and the regulation around it impact end users again requires modeling user preferences for different services. The concept of utility is a way of providing a common metric

for representing such diverse needs. This idea has its roots in economics, where it is used for a similar reason, namely, to capture an individual agent's preferences for consumption of different allocations of goods or services (Mas-Colell et al. 1995). It is a central concept in economics and also widely used in game theoretic studies to model the payoffs of players in a game.

In the literature on wireless networking, utility functions are used in several distinct ways (Berry and Johari 2013). In some work these functions can be viewed as a semantic device used to explain the design and function of the distributed algorithms needed to operate these networks. In many cases these algorithms are such that the users of the network act as if they are optimizing some underlying utility. Here the choice of utility reflects the engineering decisions made when designing the algorithms. An example of such an approach is the use of utility functions to design different congestion control algorithms (e.g., Kelly 1997). This can contrast with work in which utility is used to capture the underlying economic preferences of the users of a network. In this case, the choice of utility is not an engineering decision but rather an attempt to model the incentives of actual users and so is much closer to the use of utility in economics. One example of this is work in the broad area of network economics (Huang and Gao 2013; Maillé and Tuffin 2014).

Preferences and Utility

In economics, utility functions are often defined in terms of an underlying *preference relation* for an agent. Suppose that there is a set X of possible outcomes available in a system, and let $\mathbf{x}$ denote a specific outcome. For example, if one is dividing a unit of bandwidth among a set of agents, X would indicate all possible divisions of this bandwidth, and $\mathbf{x}$ would indicate a specific division. A *preference relation*, $\succsim$, is a binary relation used to model a user's preferences for different outcomes so that if $\mathbf{x} \succsim \mathbf{y}$, then the user finds outcome $\mathbf{x}$ at least as "good" as outcome $\mathbf{y}$.

Generally, such preference relations are assumed to be rational as defined next.

Definition 1 A preference relation for an outcome set X is *rational* if it satisfies the following two properties:

1. *Completeness*: a preference exists between any two outcomes;
2. *Transitivity* if $\mathbf{x} \succsim \mathbf{y}$ and $\mathbf{y} \succsim \mathbf{z}$, then $\mathbf{x} \succsim \mathbf{z}$ for all $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ in X .

A utility function is a way of representing a preference relation by assigning a numerical value (called the utility) to each possible outcome that is aligned with an agent's preferences.

Definition 2 A preference relation, $\succsim$, is represented by a *utility function*, $U : X \mapsto \mathbb{R}$, if $U(\mathbf{x}) \geq U(\mathbf{y})$ whenever $\mathbf{x} \succsim \mathbf{y}$.

In other words, a larger utility indicates that an outcome is more preferred by an agent. In some examples, the user's preferences and so their utility may depend only on a lower-dimensional function of the outcome; for example, in the case of resource allocation in networks, a user's utility may be a function only of the total rate allocation they receive, in which case one often writes the utility as only depending on this lower dimensional quantity.

Representing a preference relation by a utility requires that the preference relation be rational as stated next.

Proposition 1 A preference relation can be represented by a utility function only if it is rational.

If the underlying outcome set is finite, then the necessary condition in Proposition 1 is also sufficient, i.e., any rational preference relation over a finite set of outcomes can be represented via a utility. However, if the outcome set is infinite, this need not be true (but will be true under additional continuity assumptions; see e.g., Mas-Colell et al. 1995.) Conversely, it can be shown that any utility function defined over a set of outcomes gives a rational preference relation.

The assumption of rational preferences leads to an appealing mathematical framework and is widely used. However, it is not without controversy and, in particular, is not fully supported by experimental studies. This in turn has led to interest in a range of alternative ways of modeling preferences such as models for bounded rationality (Rubinstein 1998).

Comparing Utilities

Ordinal Utility

For a utility to represent a preference relation as in Definition 2, only the *relative* ranking of the utilities matters and not the *absolute* numeric values of the utility. In other words, whether an agent prefers $\mathbf{x}$ to $\mathbf{y}$ only depends if $U(\mathbf{x})$ is larger or smaller than $U(\mathbf{y})$. For this reason such a utility function is referred to as an *ordinal utility*. Note that as a consequence, the ordinal utility representing an agent's preferences is not unique. Indeed, if we set $V(\mathbf{x}) = f(U(\mathbf{x}))$ for any strictly increasing function f , then $V(\mathbf{x})$ is another valid utility function that represents the same preferences as $U(\mathbf{x})$.

Pareto Efficiency

Ordinal utility suffices for representing the preferences of one agent. However, one would often like to compare outcomes for a group of agents. The notion of Pareto efficiency (also called Pareto optimality) provides one basic way of doing this.

Consider a situation in which there are N agents impacted by the choice of outcome from the set X . Each agent i 's preferences are represented via a utility $u_i(\mathbf{x})$.

Definition 3 An outcome $\mathbf{x}$ in X is *Pareto efficient* if there is no other outcome $\mathbf{y}$ so that $U_i(\mathbf{y}) \geq U_i(\mathbf{x})$ for all agents i and this inequality is strict for at least one agent j .

In other words, for an outcome to be Pareto efficient, there must be no way to improve the outcome for one agent without making the outcome worse for other agents. Pareto efficiency provides a baseline notion of efficiency but in

many cases does not lead to a unique choice of outcomes. For example, suppose that there is a single (indivisible) good to be allocated to one of N agents. Each agent strictly prefers getting the good to not getting it; if an agent does not get the good, she is indifferent between which other agent gets it. Allocating the good to any agent is then Pareto optimal. To choose a particular one of these outcomes, it is natural to ask which agent values the good the most, i.e., which agent receives the "most utility" from the good. Without additional assumptions, an ordinal utility does not allow us to answer such questions. Since the units of utility have no meaning, it is not possible to make comparisons across agents.

When faced with multiple Pareto optimal outcomes, various *fairness* notions may be used to decide on one of them. A key distinction here is that Pareto optimality is an objective notion for characterizing a "good" outcome, while fairness is inherently subjective. Alternatively, one can make further assumptions on utility functions that facilitate comparisons. The class of quasilinear utilities described next provides one way of doing this.

Quasilinear Utilities

A common approach to facilitate comparisons across utility functions is to introduce money into the model and assume that the agents have preferences which are linear in their amount of money.

In such a setting, the set of outcomes X can be expanded to include monetary transfers, i.e., if there are N agents participating in the system, an outcome will now be given by $(\mathbf{x}, \mathbf{m})$ where $\mathbf{x}$ is again an outcome from X and $\mathbf{m} = (m_1, m_2, \dots, m_N)$ is a vector denoting the money m_i allocated to each agent i . Under a quasilinear utility model, each agent i 's net utility for this outcome are then given by

$$u_i(\mathbf{x}) + m_i,$$

where $u_i(\mathbf{x})$ indicates agent i 's value for the outcome $\mathbf{x}$. The quasilinear assumption means that we can view the valuation $u_i(\mathbf{x})$ for each agent i as being measured in monetary units; i.e.,

this represents the amount of money an agent is willing to pay for the given outcome (because of this, the quantity $u_i(\mathbf{x})$ is sometimes referred to as an agent's *willingness to pay*).

Quasilinear utilities are often assumed in the literature, in particular, in cases where market-based mechanisms are used to decide on an outcome (in which the consideration of money is natural). This assumption greatly simplifies analysis. Indeed, many key results in areas such as mechanism design or welfare economics are based on quasilinear utilities. For example, it leads to a much simpler characterization of Pareto efficient outcomes as shown in the next theorem.

Theorem 1 *Suppose all users have quasilinear utility and that any monetary allocation is allowed that satisfies $\sum_i m_i \leq 0$. Then an allocation $(\mathbf{x}^*, \mathbf{m}^*)$ is Pareto optimal if and only if it satisfies the following two conditions:*

$$\mathbf{x}^* \in \arg \max_{\mathbf{x} \in X} \sum_i U_i(\mathbf{x});$$

$$\sum_i m_i^* = 0.$$

One can view this as a situation where each agent initially has no money; agents are allowed to transfer money to other agents with no credit restrictions. The constraint $\sum_i m_i \leq 0$ means that money may leave the system but no external money can be added. The result then shows that at a Pareto optimal allocation, the sum of these transfers will be zero. Essentially, one can use monetary transfers to compensate any agent for a loss of utility that leads to an overall increase in the total utility.

Because of this result, in quasilinear environments, Pareto optimality corresponds to maximizing the sum of the valuations across the users. This sum is often referred to as the *social welfare* of the allocation. However, quasilinearity is a strong restriction on user preferences that is not supported in all scenarios. For example, this does not apply to settings in which agents exhibit a “wealth effect,” meaning that they are willing to pay more for an item when they have more money available.

Expected Utility

Utilities can also be used when there is uncertainty in the outcome. Such uncertainty is often modeled by assuming that under a given decision, outcomes in a set X occur with a given probability distribution also referred to as a *lottery*. For example, if there are only two possible outcomes, $\mathbf{x}_1$ and $\mathbf{x}_2$, then a lottery $p = (p_1, p_2)$ would indicate that outcome $\mathbf{x}_i$ occurs with probability p_i . Let $\mathcal{P}$ denote the set of all lotteries over a given outcome set X , e.g., for our two outcome example:

$$\mathcal{P} = \{(p_1, p_2) : p_i \geq 0 \text{ and } p_1 + p_2 = 1\}.$$

Agents can then be viewed as having a preferences over lotteries, modeled via a preference relation $\succsim$ over $\mathcal{P}$.

Given two lotteries p and p' over a given set of outcomes, one can combine them to form a new lottery by using lottery p with probability α and lottery p' with probability $(1 - \alpha)$. Such a compound lottery is denoted by $\alpha p + (1 - \alpha)p'$. Using this notion, the following two properties of a preference over lotteries can be defined.

Definition 4 A preference relation $\succsim$ over $\mathcal{P}$ is *continuous* if given any three lotteries p , p' , and p'' with $p \succsim p' \succsim p''$, there exists a probability α so that the lottery $\alpha p + (1 - \alpha)p''$ is equivalent to p' .

Definition 5 A preference relation $\succsim$ over $\mathcal{P}$ satisfies *independence* if for any three lotteries p , p' , and p'' , $p \succsim p'$ if and only if $tp + (1 - t)p'' \succsim tp' + (1 - t)p''$ for any probability t .

The most common way of using utility functions to represent such preferences is via an expected utility representation as defined next.

Definition 6 A preference relation $\succsim$ over the lotteries $\mathcal{P}$ on a set of outcomes X has an *expected utility representation* if there exists a utility function $u : X \mapsto \mathbb{R}$ such that if $p \succsim q$ then the expected utility under p is no smaller than that under q .

In other words, if a preference has an expected utility representation, then one only needs to define a utility function over the underlying set of outcomes and not directly on the space of lotteries over these outcomes. Lotteries are then compared by taking the expected value of this utility.

For a preference to have an expected utility representation, this relation again needs to be rational. However, this is not sufficient. A fundamental result from von Neumann and Morgenstern (1955) shows that a rational preference over a finite set of outcomes has an expected utility representation if and only if it is also continuous and independent (again this can be extended to infinite sets of outcomes with some additional technical assumptions). Such a utility function is also referred to as a *von Neumann-Morgenstern utility*. When studying decision-making under uncertainty, von Neumann-Morgenstern utilities are typically assumed. However, it should be noted that as with rational preferences, this assumption is again not supported by many experimental studies. Alternatives have been developed to better capture some characteristics of observed human behavior, such as *prospect theory* proposed in Kahneman and Tversky (1979).

Divisible Resources

In many networking problems, the underlying outcome can be viewed as a continuous nonnegative quantity x that is allocated to a given agent. For example, x could represent the amount of some resource allocated to a given agent, where we assume that this resource is perfectly divisible. In such a setting, we can view an agent's utility as a function defined on the positive real line (or some subset of it). In such contexts there are two other properties of utility functions that are natural to consider.

Monotonicity. Often one only considers utility functions that are nondecreasing in the allocation x . This means that every individual prefers more resources to less resources. A key implicit reason that monotonicity is plausible is the notion of

free disposal: that excess resources can always be “disposed of” without any cost or penalty. With free disposal, an individual is never worse off with additional resources.

Concavity. Utility functions can take any shape or form, but one important distinction is between *concave* and *convex* utility functions. Concave utility functions represent diminishing marginal returns to increasing amount of resource, while convex utility functions represent increasing marginal returns to increasing amount of resource.

In the context of network resource allocation, there is an additional interpretation of concavity that is sometimes useful. Consider two users with the same utility function $U(x)$ as a function of a scalar data rate allocation x ; these users share a single link of unit capacity. We consider two different resource allocations. In the first, we randomly allocate the entire link to one of the two users. In the second, we split the link capacity equally between the two users. Which scenario do the users prefer?

Observe that if the users are making a phone call which requires the entire link for its minimum bit rate, then the second scenario is undesirable. On the other hand, if the users are each downloading a file, then the second scenario delivers a perfectly acceptable average rate (given that both users are contending for the link). The first (resp., second) case is often modeled by assuming that U is convex (resp., concave), because the expected utility to each user in the first allocation is *higher* (resp., lower) than the expected utility to each user in the second allocation. (Both results follow by Jensen's inequality.) For this reason, concave utility functions are sometimes thought to correspond to *elastic* traffic and applications, such as file-sharing downloads, while convex utility functions are often used to model applications with *inelastic* minimum data rate requirements, such as voice calls.

Finally, from a technical standpoint, we note that concavity plays an important role when we consider maximization of utility, given the technical simplicity of concave maximization problems.

Risk Aversion

The discussion of inelastic and elastic utility functions above is closely connected to the notion of *risk aversion* in economic modeling. To see the connection, suppose $U(w)$ represents an agent's utility for w units of currency (or wealth). Suppose the individual is offered a choice between W units of wealth with certainty or a gamble that pays zero with probability $1/2$ or $2W$ with probability $1/2$.

Note that the same kind of reasoning as above reveals that if U is concave, then the individual prefers the sure thing (W units of wealth with certainty), while if U is convex, the individual prefers the gamble. For this reason we say that an individual with concave utility function is *risk averse*, while an individual with a convex utility function is *risk seeking*. It is generally accepted that individuals tend to be risk averse at higher wealth levels and risk neutral (or potentially even risk seeking) at lower wealth levels.

Cross-References

- [Efficiency and Pareto Optimality](#)
- [Fairness](#)

References

- Berry RA, Johari R (2013) Economic modeling in networking: a primer. *Found Trends Netw* 6(3):165–286
- Huang J, Gao L (2013) Wireless network pricing. *Synth Lect Commun Netw* 6(2):1–176
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decisions under risk. *Econometrica*, 47(2): 263–291
- Kelly F (1997) Charging and rate control for elastic traffic. *Trans Emerg Telecommun Technol* 8(1):33–37
- Maillé P, Tuffin B (2014) Telecommunication network economics: from theory to applications. Cambridge University Press, Cambridge
- Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic theory*. Oxford University Press, New York
- Rubinstein A (1998) *Modeling bounded rationality*. MIT Press, Cambridge
- von Neumann J, Morgenstern O (1955) *Theory of games and economic behavior*. Princeton University Press, Princeton

Price Agreement

- [Collusion](#)

Primary Synchronization Signal Sequences

- [The Primary Synchronization Sequences in LTE](#)

Principle of OFDM and Multi-carrier Modulations

Hao Li

Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada

Synonyms

[Multi-channel modulation](#); [Multi-tone modulation](#); [Orthogonal FDM](#)

Definitions

Orthogonal frequency division multiplexing (OFDM) is a transmission technique that divides the whole bandwidth into many subchannels that are closely spaced with overlapping spectra and mutually orthogonal to each other, and transmits a number of parallel low-rate data streams that are split from a higher-rate serial data stream over these orthogonal subchannels simultaneously.

Multi-carrier modulation (MCM) is a transmission technique that divides the whole bandwidth into many subchannels, without a specific orthogonality requirement, and transmits multiple data streams in parallel over the subchannels. OFDM is a special form of MCM.

Historical Background

The development of OFDM dates back to 1950s when the initial MCM system was developed for the military high-frequency (HF) radio links (Dolez et al. 1957; Mosier and Clabaugh 1958). The original MCM systems, which transmitted parallel data streams over nonoverlapped channels separated by steep-sided filters, suffered low spectrum efficiency and high implementation complexity. To overcome these challenges, a number of efforts have been dedicated to achieve MCM with overlapped channels in the 1960s. In 1961, Franco and Lachs proposed a multi-tone, code-multiplexing scheme (Franco and Lachs 1961), in which sine and cosine waves were used to generate orthogonal signals. The concept of OFDM was first introduced by Chang in 1966 (Chang 1966). Fourier transform was exploited to produce orthogonality between the subchannels over a band-limited channel. In 1967, Saltzberg extended Chang's work to complex data and proposed staggered QAM technique to reduce cross talk between adjacent channels (Saltzberg 1967). The orthogonality of subchannels in the analog OFDM system was preserved when the number of subcarriers was not large.

A digitally implemented OFDM system was first introduced by Weinstein and Ebert in 1971 (Weinstein and Ebert 1971), in which the discrete Fourier transform (DFT) was proposed to generate orthogonal subcarriers at the transmitter. This DFT-based OFDM totally eliminated bank of subcarrier oscillators at the transmitter/receiver, so as to reduce the implementation complexity. Another milestone in OFDM came in 1980, when Peled and Ruiz introduced a cyclic extension to the OFDM signal (Peled and Ruiz 1980), which is more commonly referred to as cyclic prefix (CP) today. The CP effectively turns the linear convolution of a multipath channel into circular convolution, which can eliminate inter-symbol interference (ISI) and ensure orthogonality over a time dispersive channel as long as the CP is longer than the impulse response of the channel.

OFDM gained much more popularity and attentions after the introduction of DFT and CP. Many works in the 1980s suggested the

use of OFDM for high-speed moderns and digital mobile communication systems. In 1985, OFDM was first applied in mobile wireless communications by Cimini (1985). In the late 1980s, the work began on the development of OFDM for commercial use. The first OFDM-based standard Digital Audio Broadcasting (DAB) was ratified in 1995 (DAB 1995). Since then, OFDM has been adopted in many wireless communication standards applied to wireless personal area networks (WPAN), wireless local area networks (WLAN), and wireless metropolitan networks (WMAN). In parallel to the commercial development, techniques to improve the performance of OFDM in different channel environments and diverse applications were also developed, such as time and frequency synchronization (Wang et al. 2003), peak-to-average power ratio (PAPR) reduction (Wang et al. 1999), and transmission security enhancement (Li et al. 2015).

Foundations

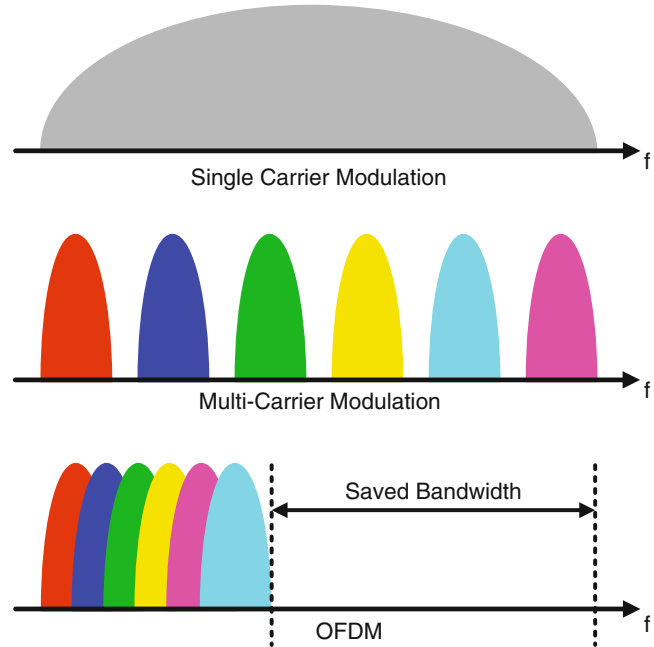
The principle of OFDM and MCM is illustrated in Fig. 1. Compared with the conventional single-carrier modulation that transmits a data stream serially over one radio frequency (RF) channel, MCM divides the RF channel into multiple subchannels and transmits multiple data streams over them in parallel. As a special form of MCM, OFDM divides the entire channel into many narrowband flat-fading subchannels that are orthogonal to each other, splits the serial data stream in parallel, and then transmits these parallel data streams simultaneously over the orthogonal subchannels with a one-to-one correspondence. A time-domain OFDM signal with N subchannels (which are also called subcarriers) can be expressed as

$$x(t) = \sum_{k=0}^{N-1} X(k)e^{j2\pi f_k t}, \quad 0 \leq t \leq T, \quad (1)$$

where $X(k)$ is the data stream transmitted on the k th subcarrier and T is the duration of an OFDM

Principle of OFDM and Multi-carrier Modulations, Fig. 1

Comparison of single-carrier modulation, multi-carrier modulation, and OFDM



signal. $f_k = k\Delta f$ is the frequency of the k th subcarrier, where Δf is the subcarrier spacing and $T\Delta f = 1$.

With a sampling interval of $T_s = \frac{T}{N}$, the sampled version of the OFDM signal $x(t)$, denoted as $x(n)$, can be given by

$$x(n) = \sum_{k=0}^{N-1} X(k) e^{j2\pi f_k \frac{nT}{N}}, \quad n = 0, 1, \dots, N-1. \quad (2)$$

Substituting $f_k = \frac{k}{T}$ into (2), the above equation can be rewritten as

$$\begin{aligned} x(n) &= \sum_{k=0}^{N-1} X(k) e^{j2\pi \frac{k}{T} \frac{nT}{N}} \\ &= \sum_{k=0}^{N-1} X(k) e^{j2\pi \frac{kn}{N}}, \quad n = 0, 1, \dots, N-1. \end{aligned} \quad (3)$$

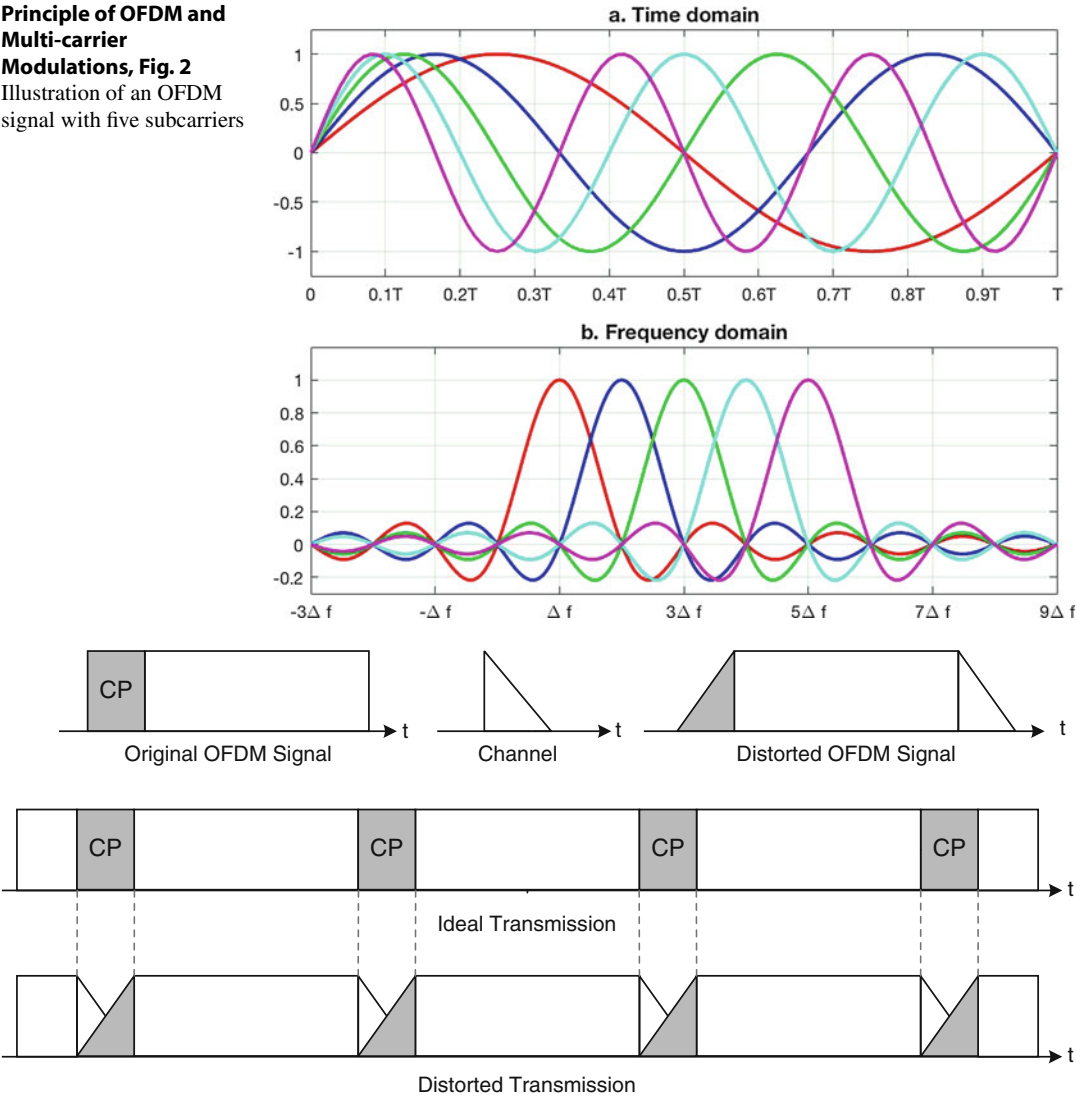
It can be found that (3) is exactly the expression of inverse discrete Fourier transform (IDFT) of $X(k)$. Therefore, the OFDM modulation process can be effectively implemented by using IDFT

or even a more efficient algorithm inverse fast Fourier transform (IFFT) of reduced computational burden. Figure 2 provides an illustration of an OFDM signal with five subcarriers, where Fig. 2a represents the construction of the time-domain OFDM signal and Fig. 2b depicts how the OFDM signal looks like in the frequency domain.

To deal with the time dispersion of wireless channels, a CP is generally employed in each OFDM signal. The basic idea of CP is to repeat the end part of a time-domain OFDM signal in its beginning to create a guard period between adjacent OFDM signals. As long as the duration of CP is longer than the maximum excess delay of the wireless channel, the ISI corrupted part of an OFDM signal stays within the guard period and can be removed later at the receiver. Hence, the impact of multipath distortion on the orthogonality among the OFDM subcarriers can be removed. Figure 3 demonstrates the principle of ISI elimination in OFDM using CP.

In addition to the spectrum efficiency achieved by the 50% overlapping between subchannels, OFDM is more robust against multipath propagation in comparison with the single-carrier system. OFDM has a larger symbol duration since the data stream is now split and transmitted in par-

Principle of OFDM and Multi-carrier Modulations, Fig. 2
 Illustration of an OFDM signal with five subcarriers



Principle of OFDM and Multi-carrier Modulations, Fig. 3 ISI elimination in OFDM using cyclic prefix

allel with a lower data rate on the orthogonal subchannels. This makes the transmission resilient to ISI and inter-frame interference. By dividing the overall frequency-selective fading channel into multiple narrowband subchannels, each subchannel is now affected individually as a flat-fading channel. Also, channel estimation and equalization can be easily applied on these flat-fading subchannels. Moreover, interference appearing on a channel could be bandwidth limited so that only data on part of the subchannels would be

affected in OFDM transmission. However, it is also noteworthy that synchronization error could destroy the orthogonality of OFDM subchannels and then disturb the transmission. OFDM thus requires highly accurate time and frequency synchronization. In addition, an OFDM signal could have a high PAPR, as it is a superposition of a large number of statistically independent symbols transmitted on all its subchannels. The signal peaks need to be mitigated in the practical implementation.

Key Applications

As an attractive transmission technique of high spectrum efficiency and robustness to channel fading, OFDM has been employed by various existing and evolving standards and products. It has been widely exploited in wired communications like digital subscriber line (DSL) and digital cable TV transmission. Also, OFDM is used as the transmission technique by powerline devices to provide high-speed local area networking over existing power lines. Meanwhile, OFDM has been adopted in major wireless communication standards, including European DAB; worldwide terrestrial digital video broadcasting like DVB-T, T-DMB, and ISDB-T; as well as 4G mobile phone standards 3GPP long-term evolution (LTE) and LTE-Advanced. Moreover, OFDM has been approved or considered by many IEEE standard working groups, such as IEEE 802.11a/g/n/ac/ax/ad/ax for WLAN, IEEE 802.15.3a for WPAN, IEEE 802.16 for WMAN, IEEE 802.20 for mobile broadband wireless access (MBWA), and IEEE 802.15.7 for visible light communication. In addition, OFDM is being investigated as one of the most promising radio transmission techniques of the 3GPP 5G wireless networks (Farhang-Boroujeny and Moradi 2016).

Cross-References

- [Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation](#)
- [Frequency-Domain Equalization in OFDM and Single-Carrier Modulation](#)
- [PAPR Reduction in OFDM Systems](#)
- [Time and Frequency Synchronization in OFDM Systems](#)

References

- Chang R (1966) Synthesis of band-limited orthogonal signals for multichannel data transmission. *Bell Syst Tech J* 45:1775–1796

- Cimini L (1985) Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing. *IEEE Trans Commun* 33:665–675
- DAB (1995) Radio broadcasting systems; digital audio broadcasting (DAB) to mobile, portable and fixed receivers. ETSI standard ETS 300 401
- Dolez M, Heald E, Martin D (1957) Binary data transmission techniques for linear systems. *Proc IRE* 45: 656–661
- Farhang-Boroujeny B, Moradi H (2016) OFDM inspired waveforms for 5G. *IEEE Commun Surv Tut* 18: 2474–2492
- Franco G, Lachs G (1961) An orthogonal coding technique for communications. *IRE Int Conv Rec* 8: 126–133
- Li H, Wang X, Chouinard JY (2015) Eavesdropping-resilient OFDM system using sorted subcarrier interleaving. *IEEE Trans Wirel Commun* 14:1155–1165
- Mosier R, Clabaugh R (1958) Kineplex, a bandwidth-efficient binary transmission system. *AIEE Trans* 76:723–728
- Peled A, Ruiz A (1980) Frequency domain data transmission using reduced computational complexity algorithms. In: *IEEE International Conference Acoustics, Speech Signal Process*, pp 964–967
- Saltzberg B (1967) Performance of an efficient parallel data transmission system. *IEEE Trans Commun Technol* 15:805–811
- Wang X, Tjhung T, Ng C (1999) Reduction of peak-to-average power ratio of OFDM system using a companding technique. *IEEE Trans Broadcast* 45: 303–307
- Wang X, Tjhung T, Wu Y, Caron B (2003) SER performance evaluation and optimization of OFDM system with residual frequency and timing offsets from imperfect synchronization. *IEEE Trans Broadcast* 49: 170–177
- Weinstein S, Ebert P (1971) Data transmission by frequency-division multiplexing using the discrete Fourier transform. *IEEE Trans Commun Technol* 19:628–634

Privacy Amplification

- [Wireless Key Establishment](#)

Privacy Attack and Defense Game

- [Privacy Game](#)

Privacy Game

Hang Shen

Department of Computer Science and
Technology, Nanjing Tech University, Nanjing,
Jiangsu Province, China

Synonyms

Game theory based privacy preservation; Privacy
attack and defense game

Definition

In general, privacy game can be defined as individual and/or cooperative behaviors between the user and the adversary (attacker) that choose a certain defense/attack strategy/action to maximize their self-interest.

Take Stackelberg privacy game (Shokri 2015; Shokri et al. 2012) between the user and the adversary as an example. The user plays first by choosing a privacy protection mechanism and committing to it by running it on his or her actual data. The follower (adversary) plays next by inferring the user's data, knowing the protection mechanism that the user has committed to. Specifically, the Stackelberg privacy game is defined as:

1. Nature selects a data (i.e., user secret that contains privacy information) for the user, where the data are selected according to a probability distribution function.
2. Given the chosen data, the user runs a protection mechanism to generate a disguised data (subject to some constraints predetermined) to release.
3. Having observed the released data, the adversary selects an estimated data by inference function. The adversary knows the internal implementation and the user's data probability distribution.

4. The adversary pays an amount (often represented by inference error equivalent to the privacy gained by the user) to the user.

In such a game, both the user and the adversary aim to maximize their payoff, i.e., the adversary tries to minimize the amount he will pay, while the user tries to maximize it. That is, they form a game-theoretic equilibrium, which is a solution for a privacy game.

Historical Background

Game theory acts as a mathematical tool used to describe and solve games (Liang and Xiao 2013). In many cases, privacy can be achieved through cooperation among entities. Users should evaluate their privacy themselves and investigate different strategies to set their privacy at their chosen level. Some researchers began to try to employ game theory – a mathematical tools for multiplayer strategic decision making – to model the interactions of agents in privacy problems. Furthermore, the theory of mechanism design (Nisan and Ronen 1999; Nisan 2007) has enabled researchers to design privacy protection mechanisms based on the analytical results (e.g., equilibrium analysis of the game). Attack and defense decisions arrived at using such game-theoretic approaches help to allocate limited resources, balance perceived risks, and take into account the underlying incentive mechanisms (Manshaei et al. 2013).

Typical Applications

According to Manshaei et al. (2013), we summarize typical applications as follows.

1. Location Privacy

Freudiger et al. (2009) first present a user-centric location privacy model to capture the evolution of location privacy for each user over time. The model considers the beliefs of users

about the tracking power of the adversary, the amount of anonymity that users obtain in the mix zones, the cost of pseudonyms, and the time of changing pseudonyms. A static game is defined where the players are mobile nodes. Each player has two strategies: Cooperate and thus change her pseudonym at mix zones, or Defect. The payoff can be calculated using the user-centric location privacy model.

2. Economics of Anonymity

Anonymity and privacy cannot be obtained by individual senders or receivers. The users should trust the infrastructure to provide protection, and other users must use the same infrastructure as well. Acquisti et al. (2003) explore the incentives of participants to offer and use anonymity services with a game-theoretic approach. They analyze the economic incentive for users to send their messages through mixnets (Chaum 1981).

3. Trust Versus Privacy

Network nodes need to disclose some private information in order to establish trust. This causes a tension between security and trust in networks. Using a game-theoretic approach, Raya et al. (2010) study this trade-off. They model the strategies of rational users that try to establish data-centric trust in the presence of rational adversaries.

4. Miscellaneous

Zhang et al. (2010) address the problem of defending against entry-exit linking attacks in Tor (The Onion Routing) network using a game-theoretic approach. They formalize the problem as a repeated noncooperative game between the defender and the adversary. They extract three design principles, namely, stratified path selection, bandwidth order selection, and adaptive exit selection, by using the results of the game between attacker and defender. Using these results, they develop gPath, a path selection algorithm that integrates all three principles to

significantly reduce the success probability of linking attacks in the Tor network.

Cross-References

► Game Theory

References

- Acquisti A, Dingledine R, Syverson P (2003) On the economics of anonymity. In: *Proceedings of the Financial Cryptography (FC), Lecture Notes in Computer Science*, Springer, vol 2742. pp 84–102
- Chaum D (1981) Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun ACM* 24(2):84–88
- Freudiger J, Manshaei MH, Hubaux J-P, Parkes DC (2009) On non-cooperative location privacy: a game-theoretic analysis. In: *Proceedings of the ACM conference on computer and communications security (CCS)*, ACM, pp 324–337
- Liang X, Xiao Y (2013) Game theory for network security. *IEEE Commun Surv Tutor* 15(1):472–486
- Manshaei MH, Zhu Q, Alpcan T, Başar T, Hubaux JP (2013) Game theory meets network security and privacy. *ACM Comput Surv* 45(3):25
- Nisan N (ed) (2007) *Introduction to mechanism design (for computer scientists)*. In: *Proceedings of the algorithmic game theory*. Cambridge University Press, Cambridge, UK, pp 209–242
- Nisan N, Ronen A (1999) Algorithmic mechanism design. In: *Proceedings of the 31 annual ACM symposium on theory of computing*, ACM, pp 129–140
- Raya M, Shokri R, Hubaux J-P (2010) On the tradeoff between trust and privacy in wireless ad hoc networks. In: *Proceedings of the ACM conference on wireless network security (WiSec)*, ACM, pp 75–80
- Shokri R (2015) Privacy games: optimal user-centric data obfuscation. *Proc Privacy Enhanc Technol* 2015(2):299–315
- Shokri R, Theodorakopoulos G, Troncoso C, Hubaux JP, Le Boudec JY (2012) Protecting location privacy: optimal strategy against localization attacks. In: *Proceedings of the ACM conference on computer and communications security (CCS)*, ACM, pp 617–627
- Zhang N, Yu W, Fu X, Das SK (2010) gPath: a game-theoretic path selection algorithm to protect Tor's anonymity. In: *Proceedings of the 1st conference on decision and game theory for security (GameSec)*, *Lecture Notes in Computer Science*, Springer, vol 6442. pp 58–72

Privacy Preservation

- Long-Term Quality of Protection
- Privacy Preservation for Location-Based Services

Privacy Preservation for Location-Based Services

Qingqi Pei and Lichuan Ma
Xidian University, Xi'an, China

Synonyms

Location based services; Privacy preservation

Definitions

The Location-Based Services (LBS) are information, alerts, or entertainment services that are provided to users according to their geographical and temporal information. Examples of LBS include location-based traffic reports, store finder, advertisements, and entertainments.

Historical Background

With the popularity of location-detection devices (e.g., handled devices equipped with GPS), their holders benefit a lot from LBS applications by receiving specific and useful information based on their location data. As a result, LBSs gain a dramatic development recently and various LBS applications have appeared. According to Shengling et al. (2018), LBSs can be classified into two categories, that is, elementary services and derivative services. Two typical examples of elementary services are navigation and search of Points of Interest (POIs). In navigation, positioning systems embedded in vehicles or handled devices can provide the location data

of their users, and these data are utilized by LBS applications to offer guidance for the users. This kind of services is the foundation of the main functions of Google Maps, Baidu Map, and AutoNavi Map. With respect to the search of POIs, a user can be recommended with the places nearby that he is interested in. Typical applications of such kind of services include Yelp, AroundMe, and Dianping. Atop these elementary services, various derivative ones, like mobile commerce, location-based games, and solomo applications, have come up to provided more complicated services. Besides these two kinds of applications, the location information can be also utilized to construct session keys for private communications (Zi et al. 2017).

Obviously, LBSs have brought much convenience to people's daily life. However, a lot of sensitive information can be disclosed by the process of collecting and analyzing the locations and queries of the users. For example, when a user queries the LBS server for the nearest restaurant from him, the server can obtain the accurate location of this user at this time. If the user queries the server consequently for different services, the trajectory of this user can be inferred by the server. This trajectory might be utilized by the server to send specific advertisements to this user that makes him bored. Situations become much worse when the location information of the users is obtained by the thieves as this kind of information indicates whether this user is at home. As for the queries, they might expose some private information of the user, like his behavior patterns or preferences. In this sense, how well the privacy of the users can be preserved determines the mass market share of LBS.

Foundations

Related to privacy preservation for LBS users, a great number of works focus on how to preserve the location privacy. These works can be categorized to three types, that is, k -anonymity,

cryptography-based oblivious data search, and differential privacy-based approaches.

- ***k*-anonymity.** This is the most popular method to preserve the location privacy of LBS users. The basic idea is that the location perturbation of a certain user is achieved with other $k - 1$ users. In this manner, *k*-anonymity can make the probability of exposing a host user's location less than $1/k$. The direct application of *k*-anonymity is proposed in Haibo and Jianliang (2009) where the proximity information of a certain LBS user is used to construct a cluster with other $k - 1$ users in a peer-to-peer manner. In this way, the cloaking area is the one that is covered with the newly constructed cluster. As the method in Haibo and Jianliang (2009) is time-consuming, a real-time location anonymization algorithm is presented in Joseph and Romit Roy (2009) by introducing the concept of prediction privacy. Since it is very possible that the adversary can have side information about the user's real location, the authors of (Ben et al. 2014) design an entropy-based measure to achieve *k*-anonymity.

- **Cryptography-based oblivious data search.** In Roman et al. (2015), identity-based encryption and hash encryption are combined to construct a private information retrieval method to secure the location privacy of LBS users. Since the somewhat homomorphic encryption (SHE) schemes become much easier to be implemented and run much faster, another kind of method is to encrypt the location data under SHE schemes. The nearest services can be computed based on the encrypted data (Anh et al. 2017). This is like to find *k* nearest neighbors over encrypted data in Wai Kit et al. (2009) and no other parties except the user himself can obtain the original location data.

- **Differential privacy-based approaches.** The location privacy of the differential privacy-based approaches is achieved via adding noises to original location information. In this way, geo-indistinguishability with no loss of LBS quality is able to be satisfied (Miguel et al. 2013). Usually, these approaches apply Laplacine noise to perturb a user's location, which is the

most typical way of realizing differential privacy (Shengling et al. 2018).

Apart from location information, the queries also contain important but sensitive information about the LBS users. However, the works address query privacy are much fewer compared with the ones aiming to preserve location privacy.

To preserve the query privacy, related methods are classified into identity hiding and content hiding. Hiding identities of the LBS users means that it is impossible to link any query to a certain user. In Bugra and Ling (2008), a personalized *k*-anonymity model is proposed to guarantee the specialized anonymity level that the users desire where the spatial and temporal tolerance of a cloaking area is taken into consideration. Because the model in Bugra and Ling (2008) relies on a fully trusted anonymizer, there is a great risk that all the identities might be disclosed if the anonymizer is compromised by attackers. To address this issue, a distributed algorithm to achieve *k*-anonymity in a peer-to-peer manner is proposed in Chi-Yin et al. (2011). As stated in Lichuan et al. (2018), only hiding the identities is fragile to tracing attack by some side information. This leads to the necessity to hide the content of queries in the meanwhile. When hiding the content of the query, it means that the attributes of the queries should be hidden. One direct way is to generate several dummy queries to be sent to the LBS server at the same time (Aniket et al. 2011). By this manner, the actual query is undistinguishable from all the other queries. The main drawback of this method is introducing too much communications overhead. Also, some cryptographic techniques are also proposed. By utilizing oblivious transfer to map the private location of the user to a public cell, a query-hiding framework is proposed in Russell et al. (2013).

Future Directions

Although different countermeasures have been proposed to preserve the privacy of LBS users, the following open issues are still needed to be further explored.

• **Designing lightweight privacy preserving methods:** As smart handled devices with limited energy are prevalent nowadays, it is required to work out lightweight privacy preserving methods for LBS users. This is because additional computation and communication overhead would be introduced in the process of preserving both location and query privacy. The situation becomes much worse when complex cryptographic approaches are utilized. In order to save the energy of batteries, it is very possible for LBSs users to give up applying these energy-consuming methods.

• **Preserving location and query privacy simultaneously:** Through the analysis in Foundations, most existing works address either location privacy or query privacy. Actually, these two categories are not independent. Only focusing on one aspect means the other one would be the side information to infer the private information of LBS users. Thus, there is an urgent need to work out new methods to preserve location and query privacy simultaneously.

• **Detecting the misuse of LBSs when preserving privacy beforehand:** Preserving privacy means that it is impossible to link a location or a query to a certain user. Motivated by financial interests, some users might misuse LBSs and only preserving privacy encourages these malicious users to keep their misuse without being detected. To deal with the conflicting issues of preserving privacy and detecting misuse, new privacy preserving methods that are capable of privately proving the reliability of LBS users should be put forward.

Cross-References

- [IoT Security](#)
- [Location Based Services](#)
- [Mobile Security and Privacy](#)
- [Network Privacy Surveillance and Privacy Protection](#)
- [Privacy-Preserving Data Collection](#)

References

- Anh P, Italo D, Guillaume E, Juan T, Kevin H, Jean-Pierre H (2017) ORide: a privacy-preserving yet accountable ride-hailing service. In: The 26th USENIX security symposium, pp 1235–1252
- Aniket P, Nan Z, Xinwen F, Hyeong-Ah C, Suresh S, Wei Z (2011) Protection of query privacy for continuous location based services. In: IEEE conference on computer communications, pp 1710–1718
- Ben N, Qinghua L, Xiaoyan Z, Guohong C, L H (2014) Achieving k-anonymity in privacy-aware location-based services. In: IEEE conference on computer communications, pp 754–762
- Bugra G, Ling L (2008) Protecting location privacy with personalized k-anonymity: architecture and algorithms. *IEEE Trans Mob Comput* 7:1–18
- Chi-Yin C, Mohamed FM, Xuan L (2011) Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments. *GeoInformatica* 15: 351–380
- Haibo H, Jianliang X (2009) Non-exposure location anonymity. In: the 25th IEEE international conference on data engineering, pp 1120–1131
- Joseph M, Romit Roy C (2009) Hiding stars with fireworks: location privacy through camouflage. In: the 15th ACM annual international conference on mobile computing and networking, pp 345–356
- Lichuan M, Xuefeng L, Qingqi P, Yong X (2018) Privacy-preserving reputation management for edge computing enhanced mobile crowdsensing. *IEEE Trans Serv Comput*, Early Access, 1–14. <https://doi.org/10.1109/TSC.2018.2825986>
- Miguel E A, Nicolas E B, Konstantinos C, Catuscia P (2013) Geo-indistinguishability: differential privacy for location-based systems. In: ACM SIGSAC conference on computer and communications security, pp 901–914
- Roman S, Chi-Yin C, Qiong H, Duncan SW (2015) User-defined privacy grid system for continuous location-based services. *IEEE Trans Mob Comp* 14:2158–2172
- Russell P, Md Golam K, Xun Y, Elisa B (2013) Privacy-preserving and content-protecting location based queries. *IEEE Trans Knowl Data Eng* 26:1200–1210
- Shengling W, Qin H, Yunchuan S, Jianhui H (2018) Privacy preservation in location-based services. *IEEE Commun Mag* 56:134–140
- Wai Kit W, David Wai-lok C, Ben K, Nikos M (2009) Secure kNN computation on encrypted databases. In: ACM SIGMOD international conference on management of data, pp 139–152
- Zi L, Qingqi P, Ian M, Yao L, Haojin Z (2017) Secret key establishment via RSS trajectory matching between wearable devices. *IEEE Trans Inf Forensic Secur* 13:802–817

Privacy-Preserving Data Aggregation for Smart Grids

Chengcheng Zhao¹, Jianping He², and Lin Cai^{3,4}

¹College of Control Science and Engineering, Zhejiang University, Hangzhou, China

²The Department of Automation, Shanghai Jiao Tong University, and the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China

³The Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

⁴E&CE Department, Illinois Institute of Technology, Chicago, IL, USA

Synonyms

Data gathering and aggregation with privacy preservation in smart grids; Privacy-preserving data integration for the modernized power grid

Definitions

Privacy-preserving data aggregation (PPDA) in smart grids is the compilation of the local smart meter measurements with the intent to prepare aggregated power consumption for the control and optimization in the operation center under the privacy guarantees. PPDA in smart grids can also be interpreted as the process where raw data of smart meters is gathered and expressed in a summary form for statistical analysis without the leakage of each consumer's sensitive information.

Historical Background

Integrated with demand response, renewable energy resources, information technologies, and etc., the smart grid has become a key technology to upgrade and modernize the traditional power

grid. The terminology named smart grids was defined by the Energy Independence and Security Act in 2007, and it has received much attention from governments, industries, and researchers. With numerous smart meters and advanced monitoring systems, data aggregation is becoming a significant step to achieve efficient, stable, and secure operation of power systems. For example, the operation center needs to collect electricity consumption information for efficient power dispatch through the aggregation of smart meter measurements from end users. Meanwhile, local aggregators may interact with neighbors to obtain the aggregated information in a distributed way, which is more promising for the lightweight computation and communication, and the privacy concern. Many results have shown that the privacy of end users can be disclosed through smart meter readings. In 1992, Hart et al. investigated the power consumption identification of each household electric appliance by eye inspection of readings (Hart 1992). Then, a sequence of nonintrusive appliance load monitoring analysis technologies based on aggregated measurements in smart grids were studied (Zeifman and Roth 2011). The Website bwired.nl created by a Dutch student also shows the sensitive data disclosure of the consumer through the smart meter measurements. By exploiting the sensitive data, the privacy of the consumer, including preferences, habits, behaviors, and even beliefs, can be obtained. As a result, property and mental losses are caused to the customer. For instance, the theft may be able to sneak into an empty house or the attacker can use the private information to hack into the computers. These privacy concerns impede the development and implementation of the smart grid. For instance, due to the privacy rights, the Dutch Parliament turned down the bill for smart grid deployment in 2009. In order to preserve privacy in data aggregation, in past years, different solutions have been investigated (Garcia and Jacobs 2010; Rial and Danezis 2011). On the one hand, the investigation on the centralized PPDA has been done, where the local load perturbation based on

the storage or controllable load and the secure data transmission based on cryptographic theory were studied (Sun et al. 2018). On the other hand, much attention has been paid to distributed PPDA in smart grids, such as the privacy-preserving consensus-based energy management protocol (Zhao et al. 2017).

Foundations

Originally, aggregation is developed to deal with the computation overhead and resource allocation issues caused by the limited storage and complex analysis for the bulk quantities of measurement data. In the smart grid, PPDA aims to analyze the statistical information of the huge volumes of smart meter measurement data, including the sum of the power consumption of end users, the average or maximum (minimum) expending, and etc., while preserving the sensitive information of the end users. From the perspective of end users, there are two potential approaches of preserving privacy in the data aggregation. The first approach is the local data perturbation method where the data is perturbed before sending to the utility company. Storage devices or controllable loads, i.e., the local rechargeable battery, electric vehicles, heating, ventilating, and air conditioning (HVAC) systems, etc., are utilized to obscure the actual household energy consumption that is measured and reported by the smart meter (Sun et al. 2018). These solutions are customer-oriented, which are more desirable to accommodate individual privacy preferences. Existing schemes are described as below:

- **Battery-based load hiding (BLH) schemes:** BLH methods enclose the electricity utilization variations through regulating the charging and discharging process of the battery. The household load profile variances can be adjusted with a rechargeable battery to keep a constant output. Differential privacy and information leakage are applied to characterize the privacy-preserving degree. Such methods can be very costly to implement.
- **Load-based load hiding (LLH) schemes:** LLH methods achieve the load masking by shifting

the user's controllable loads. Under the proposed schemes, either the energy consumption is flatten over time or artificial electricity signatures are injected to hide the original load profile. Available options for the controllable loads are very limited for LLH schemes.

- **BLH-LLH schemes:** BLH-LLH methods combine BLH and LLH by exploiting both of household controllable loads and local batteries and thus to render the power consumption preservation more practical in the implementation stage.

The other approach is to preserve the privacy during the data aggregation process, which happens after the smart metering readings are broadcast. We introduce commonly used methods as follows:

- **Data anonymization:** data anonymization methods remove the attribute information from the meter readings to obscure the correspondence between smart meters and customers. However, these methods may fail when the side information of users is attainable.
- **Cryptographic-based computation (CBC) methods:** CBC schemes apply the homomorphic encryption and secret sharing, masking and brute forcing, modified homomorphic encryption, and so on, so that the aggregator cannot know the sensitive information in the measurements, while the aggregated results can be decrypted by the utility company. These methods are limited by their required computation or communication complexity, and the performance depends on specific model design with respect to the privacy protection.
- **Random noise adding (RNA) methods:** RNS schemes hide the power consumption by adding random noise to the real data. For example, differential privacy has been a hot research topic and applied in practical data integration process. It aims to maximize the accuracy of queries from statistical databases while maintaining indistinguishability of its transcripts. As a result, the utility company is able to protect the users' sensitive information

at the cost of obtaining less accurate statistical data.

Note that the above methods mainly focus on the centralized PPDA, which cannot be directly applied to the distributed cases. Under distributed PPDA, all participants interact with their neighbors to obtain the aggregated global information, while the privacy of each participant is preserved. The local load perturbation method requires that all local consumers have enough storage devices or controllable loads to hide the load profile. Data anonymization may be too complex to realize the distributed PPDA. CBC technologies may cause much communication and computation cost due to intensive interactions when the network scales up. RNA methods may be applied in the distributed PPDA, but how to define the privacy degree with partial information and how to guarantee the required accuracy with the privacy preservation guarantees are different and interesting problems. Plentiful distributed privacy-preserving schemes have been proposed for typical distributed data aggregation methods, e.g., consensus, distributed optimization, and data selection, which are introduced as follows:

- Privacy-preserving maximum consensus (PPMC) (Duan et al. 2015; Wang et al. 2017): To preserve the initial state of maximum consensus while maintaining the finite time convergence, the PPMC method was proposed in 2015, where all nodes independently generate and transmit random numbers before sending out initial states. Then, under the assumption that an adversary can attain all transmitted states, a differentially private maximum consensus (DPMC) algorithm was developed, where nodes add Laplacian noises to initial states for local interactions.
- Privacy-preserving average consensus (PPAC) (He et al. 2018a, 2016): PPAC is to guarantee average consensus while protecting the initial state of all nodes from being inferred by the internal or external attacker.
 - *Exact PPAC*: An exact average consensus with privacy preservation guarantees was achieved by adding the zero-sum decreasing noise.
 - *Differential PPAC*: To guarantee differential privacy, a typical method is adding random noise to the original data for data release. The basic conditions of differential privacy with the general random noise adding mechanism were investigated.
 - *Secure PPAC*: To mitigate the pollution from dishonest nodes, an enhanced secure consensus-based data aggregation algorithm was proposed that allows neighbors to detect dishonest nodes and guarantees an error bound for the case with undetectable dishonest nodes.
- Privacy-preserving distributed optimization (PPDO) (Nozari et al. 2018): PPDO obscures the private local function by noisy message exchange or local function perturbation, under which multiple agents are still able to cooperatively minimize the sum or average of their local cost functions with/without constraints to some extent.
 - Local message perturbation (LMP): LMP hides the sensitive local cost function by adding noise to the messages which are broadcast to neighbors while having the optimality guarantee, for example, differentially private distributed optimization algorithms.
 - Local function perturbation (LFP): LFP protects the private local cost function by obscuring it, where the objective function is perturbed by noise, while the accuracy of the global optimizer has an analytical bound.
- Privacy-preserving distributed selection (PPDS) (Liu and Chen 2018): PPDS hides the local statistic information of distributed data owner from neighbors or the eavesdropper, e.g., mean and sum, while the required states can still be selected in a distributed way, e.g., the median.
- Distributed privacy-preserving estimation theory (He and Cai 2017; He et al. 2018b): To preserve data privacy, a node may add random noise to its original data for information exchange at each iteration. Nevertheless, an eavesdropping node is able to estimate the

other's original data depended on its received information. It is worth mentioning the following results:

- The estimation accuracy and data privacy are measured in terms of (ϵ, δ) -data privacy, defined as the probability of ϵ -accurate estimate (the difference between an estimation and the original data that is within ϵ is no larger than δ).
- A theoretical framework which achieves the optimal distributed estimation is developed, where the agents optimize the estimation of neighbors' original data using the local information.
- The disclosure probability is derived under the optimal estimation. These theoretical results have been successfully applied in distributed energy management for smart grids (Zhao et al. 2017).

Key Applications

PPDA involves in the data processing for the statistical analysis for huge volumes of measurement data in a geographically distributed or logically distributed network. Typical application scenarios include smart meters data aggregation in smart grids, time synchronization for wireless sensor networks, and efficient resource allocation in end-to-end networks.

Cross-References

► [Sparse Signal Processing](#)

References

- Duan X, He J, Cheng P, Mo Y, Chen J (2015) Privacy preserving maximum consensus. In: Decision and control (CDC), 2015 IEEE 54th annual conference on, IEEE, Osaka, Japan pp 4517–4522
- Garcia FD, Jacobs B (2010) Privacy-friendly energy-metering via homomorphic encryption. In: International workshop on security and trust management, Springer, pp 226–238

- Hart GW (1992) Nonintrusive appliance load monitoring. *Proc IEEE* 80(12):1870–1891
- He J, Cai L (2017) Differential private noise adding mechanism: basic conditions and its application. In: American control conference (ACC), 2017, IEEE, Seattle, WA, USA pp 1673–1678
- He J, Cai L, Zhao C, Cheng P, Guan X (2016) Privacy-preserving average consensus: privacy analysis and optimal algorithm design. *arXiv preprint arXiv:160906368*
- He J, Cai L, Cheng P, Pan J, Shi L (2018a) Distributed privacy-preserving data aggregation against dishonest nodes in network systems. *IEEE Internet Things J.* <https://doi.org/10.1109/JIOT.2018.2834544>
- He J, Cai L, Guan X (2018b) Preserving data-privacy with added noises: optimal estimation and privacy analysis. *IEEE Trans Inf Theory.* <https://doi.org/10.1109/TIT.2018.2842221>
- Liu H, Chen J (2018) Distributed privacy-aware fast selection algorithm for large-scale data. *IEEE Trans Parallel Distrib Syst* 29(2):365–376
- Nozari E, Tallapragada P, Cortés J (2018) Differentially private distributed convex optimization via functional perturbation. *IEEE Trans Control Netw Syst* 5(1):395–408
- Rial A, Danezis G (2011) Privacy-preserving smart metering. In: Proceedings of the 10th annual ACM workshop on privacy in the electronic society, ACM, Berlin, Germany pp 49–60
- Sun Y, Lampe L, Wong VW (2018) Smart meter privacy: exploiting the potential of household energy storage units. *IEEE Internet Things J* 5(1):69–78
- Wang X, He J, Cheng P, Chen J (2017) Differentially private maximum consensus. *IFAC-PapersOnLine*, Toulouse, France 50(1):9509–9514
- Zeifman M, Roth K (2011) Nonintrusive appliance load monitoring: review and outlook. *IEEE Trans Consum Electron* 57(1):76–84
- Zhao C, He J, Cheng P, Chen J (2017) Privacy-preserving consensus-based energy management in smart grid. In: Power & energy society general meeting, 2017 IEEE, IEEE, Chicago, IL, USA pp 1–5

Privacy-Preserving Data Collection

Zhan Qin

Texas University at San Antonio, San Antonio, TX, USA

Synonyms

[Secure data collection](#)

Definitions

Privacy-preserving data collection system collects sensitive data from data contributors and detects meaningful general information by gathering data while protecting data contributors' privacy.

Background

Nowadays, with the advance of big data analytics, organizations have become increasingly interested in collecting and analyzing user data. For example, web browsers and mobile apps often collect system logs and usage patterns as a means to guide the development of future versions; crowdsourcing platforms, such as Mechanical Turk, also provide a convenient way to collect information from contributors. However, the collection of user data could incur significant privacy risks, as demonstrated in several past incidences, e.g., (Hansell 2006), where accidental leakage of sensitive data led to public outrage, reputation damage, and legal actions against the data collector. The need of a privacy-preserving way to collect data from users has been recognized by both industry and academia.

Local differential privacy (LDP) is the state-of-the-art approach for *privacy-preserving data collection*, which has been implemented in the Google Chrome browser (Erlingsson et al. 2014). Its main idea is to ensure that the data collector (i) never collects or possesses the exact values of any personal data and yet (ii) would still be able to derive general statistics about the users. In particular, in LDP, each user locally perturbs her data under differential privacy (Dwork 2008), which is a strong and rigorous notion of privacy that provides plausible deniability; then, the user sends the perturbed data to the collector. As such, LDP protects both the users (against privacy risks) and the data collector itself (against damages caused by potential privacy breaches).

A series of works study the problem of frequency estimation following local differential privacy. After Warner's first study on *randomized*

response (Warner 1965), many works have explored a variety of perturbation mechanisms for achieving theoretical optimal utility. In (Kasiviswanathan et al. 2011), the local privacy model is first formalized. To handle the practical issues for private distribution estimation, RAPPOR (Erlingsson et al. 2014) is proposed to generalize Warner's randomizer from binary alphabets to k-ary alphabets.

Foundations

Privacy-preserving data collection system with local differential privacy (LDP) (Kasiviswanathan et al. 2011) is a data collection framework based on ϵ -differential privacy (Dwork 2008). Under this framework, each data contributor (e.g., a user of a website) locally perturbs her own data using a randomized mechanism that satisfies ϵ -differential privacy, before sending the noisy version of her data to a data collector. Specifically, a randomized mechanism M satisfies ϵ -differential privacy, if and only if for any two neighbor databases D and D' (explained shortly) that differ in exactly one record, and any possible output s of M , we have the following inequality:

$$\frac{Pr[M(D) = s]}{Pr[M(D') = s]} \leq e^\epsilon$$

Intuitively, given an output s of mechanism M , an adversary cannot infer with high confidence (controlled by ϵ) whether the input database is D or its neighbor D' , which provides plausible deniability for individuals involved in the sensitive database. Here, ϵ is a system parameter called the privacy budget that controls the strength of privacy protection. A smaller ϵ signifies stronger privacy protection, since the adversary has lower confidence when it tries to distinguish between D and D' .

The concept of differential privacy was originally proposed for the setting where a trusted data curator, who possesses a database containing the exact data records from multiple individuals,

publishes perturbed statistics derived from the database using a randomized mechanism. Hence, the definition of differential privacy involves the notion of “neighbor databases.” In contrast, in LDP, there is no trusted data curator; instead, each individual perturbs her own data record under ε -differential privacy. In this situation, D and D' are two singleton databases, each containing exactly one record. Since two neighbor databases differ by exactly one record, essentially D and D' represent two arbitrary records. In the problem studied in this paper (described in section “[Background](#)”), every record is a set of items, and each of D and D' represents such an item set.

A privacy-preserving data collection solution includes both (i) a user-side data perturbation mechanism and (ii) an algorithm executed by the data collector for computing statistical information from the noisy data received from the users. The former one’s main requirement is to satisfy differential privacy. Theoretically, every differentially private mechanism can be potentially applied to the LDP setting to be run at each user. In reality, however, most mechanisms for enforcing differential privacy are designed to run at a central data curator who has access to the exact records of all users, with the goal of publishing perturbed statistics based on these records. In LDP, neither assumption holds, since the privacy protection algorithm is run by individual users who do not have access to other user’s data, and that the information to be released is the perturbed data, not analysis results. Consequently, although the fundamental mechanisms for differential privacy such as the Laplace mechanism and the geometric mechanism can be applied to LDP, more sophisticated ones, e.g., for the frequent itemset mining, cannot be applied to LDP as they require global knowledge of all users’ data. Further, perturbing the data at the user side is only part of the solution; traditional differentially private mechanisms do not address the other important problem in LDP, i.e., computing statistics at the data collector based on individuals’ noisy data.

To understand the privacy-preserving data collection system with local differential privacy, an example of randomized response mechanism

achieves local differential privacy is shown below. Specifically, randomized response asks each user a sensitive question whose answer can be either yes or no, e.g., “Are you HIV positive?” The goals are (i) that each user answers the question with plausible deniability and (ii) that the data collector can compute an unbiased estimate of the percentage of users whose answer is “yes” (resp. “no”). To do so, each user flips a coin before answering the question. If the coin turns head, the user provides her true answer (let’s say it is “yes”); otherwise, she reports the opposite of her true answer (i.e., “no”). To adapt RR to satisfy LDP, we make the coin biased, with a probability p (resp. $1 - p$) to turn head (resp. tail). It has been proven (e.g., in Erlingsson et al. 2014) that randomized response satisfies ε -differential privacy with the following value of p :

$$p = \frac{e^\varepsilon}{1 + e^\varepsilon}$$

After that, the data collector in randomized response estimates the percentage of users with an “yes” answer. If the data collector simply outputs the percentage of “yes” among the noisy answers (denoted by c), the results would be biased, due to the random perturbations performed at the users. To correct this, the data collector reports the following adjusted estimate:

$$c' = c \times c_\varepsilon, \text{ where } c_\varepsilon = \frac{1}{1 - 2p}$$

Randomized response is limited to the case where each user answers a binary question. Nevertheless, it is still a fundamental building block for even more sophisticated LDP solutions.

Cross-References

- [Private Truth Discovery in Cloud-Empowered Crowdsensing Systems](#)
- [Privacy-Preserving Healthcare Service](#)

References

- Dwork C (2008) Differential privacy: a survey of results. In: *Theory and applications of models of computation*, pp 1–19
- Erlingsson U et al (2014) RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: *CCS*, pp 1054–1067
- Hansell S (2006) Aol removes search data on vast group of web users. *New York Times* 8:C4
- Kasiviswanathan SP et al (2011) What can we learn privately? *SICOMP* 40(3):793–826
- Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. *J ASA* 60(309):63–69

Privacy-Preserving Data Integration for the Modernized Power Grid

- [Privacy-Preserving Data Aggregation for Smart Grids](#)

Privacy-Preserving Healthcare Service

- [Encrypted Search for Medical Data](#)

Privacy-Preserving Truth Inference

- [Private Truth Discovery in Cloud-Empowered Crowdsensing Systems](#)

Private Truth Discovery in Cloud-Empowered Crowdsensing Systems

Yifeng Zheng and Cong Wang
Department of Computer Science, City
University of Hong Kong, Hong Kong, China

Synonyms

[Crowdsourcing](#); [Cloud computing](#); [Mobile crowdsensing](#); [Privacy-preserving truth inference](#)

Definitions

Private truth discovery in cloud-assisted crowdsensing systems aims to mine trustworthy information from noisy sensory data collected from different mobile users in cloud-empowered crowdsensing applications while enforcing the provision of data privacy.

Historical Background

Crowdsensing is an emerging sensing paradigm which enables cost-effective sensory data collection from ubiquitous mobile devices (Ganti et al. 2011). It has found a wide spectrum of practical applications like health monitoring, indoor floor-plan reconstruction, environmental monitoring, and smart transportation. In the real practice, crowdsensing systems are usually deployed on top of the cloud to collect and process sensory data, due to its well-known advantages like scalability, economical cost, and on-demand services. As the sensory data collected from different mobile users are usually unreliable (possibly due to the variation in sensor quality, ambient noise, skill level, etc.), it is of immense importance to properly aggregate the sensory data so as to obtain trustworthy information. Toward this problem, the technique of truth discovery has been recently proposed, which is able to infer individual users' reliability information directly from the collected sensory data and then infer the truthful information via reliability-aware aggregation (Li et al. 2015).

Despite the usefulness, practically applying truth discovery in cloud-empowered crowdsensing systems faces critical privacy barriers that should be overcome satisfactorily. Firstly, the sensory data collected from the mobile users may reveal sensitive information to the cloud service provider such as health conditions, locations, personal interests, and more (Ganti et al. 2011). More seriously, the cloud service provider might also misuse the collected data for making profit or even suffer from data breaches incidents, which have been shown to be common in practice (Zheng et al. 2017a, c). Secondly,

the reliability degrees of individual users inferred from the sensory data also demand delicate protection, as it may be potentially exploited to infer sensitive information like education level, age, and profession (Miao et al. 2015). Therefore, how to preserve data privacy while maintaining the functionality of truth discovery is a critical challenge that should be carefully tackled in cloud-empowered crowdsensing systems. Toward this problem, a line of research work (Miao et al. 2015; Xu et al. 2017; Miao et al. 2017; Zheng et al. 2017b, 2018; Tang et al. 2018; Cai et al. 2018) has been presented in the literature.

Truth Discovery Framework

The most representative truth discovery framework in the plaintext domain is the CRH framework proposed by Li et al. (2016), which is also the foundation for most of existing privacy-preserving truth discovery designs. The CRH framework for truth discovery in mobile crowdsensing applications is introduced as follows. Let $\mathcal{N} = \{1, 2, \dots, N\}$ denote the set of sensing tasks and $\mathcal{U} = \{1, 2, \dots, M\}$ denote the set of users. Besides, let y_n^i denote the sensing value for the sensing task $n \in \mathcal{N}$ from user $i \in \mathcal{U}$ and w_i denote the reliability degree, i.e., weight, of user i . The CRH truth discovery framework aims to produce the estimated ground truth y_n^* for each sensing task $n \in \mathcal{N}$. There are two iterative procedures to be conducted: (1) weight update and (2) truth update.

- (1) **Weight update.** This procedure is to compute a weight w_i for user $i \in \mathcal{U}$, given that the estimated ground truth y_n^* for each sensing task n is fixed. A user will be assigned a weight with a high value if his/her sensing values are close to the estimated ground truths. The weight w_i for user i is computed as

$$w_i = f \left(\sum_{n=1}^N d(y_n^i, y_n^*) \right), \quad (1)$$

where f is a monotonically decreasing function and $d(\cdot)$ is a distance function measuring the difference between users' sensing values and the estimated ground truths. In more detail, w_i is computed via.

$$w_i = \log_2 \left(\frac{\sum_{i=1}^M \sum_{n=1}^N d(y_n^i, y_n^*)}{\sum_{n=1}^N d(y_n^i, y_n^*)} \right). \quad (2)$$

Meanwhile, the following normalized squared distance function is commonly used as the distance function d .

$$d(y_n^i, y_n^*) = \frac{(y_n^i - y_n^*)^2}{\sigma_n}, \quad (3)$$

where σ_n is the standard deviation of all sensing values for sensing task n .

- (2) **Truth update.** This procedure is to estimate the ground truth y_n^* for each sensing task $n \in \mathcal{N}$, given that each user weight w_i is fixed. Specifically, the estimated ground truth y_n^* for the sensing task n is computed via weighted aggregation of users' sensing values, i.e.,

$$y_n^* \leftarrow \frac{\sum_{i=1}^M w_i y_n^i}{\sum_{i=1}^M w_i}. \quad (4)$$

Advances on Privacy-Preserving Truth Discovery

In the literature, some research efforts have been recently presented to enable privacy-preserving truth discovery for cloud-empowered crowdsensing systems. Here a brief overview of these works is provided.

The initial research efforts were from Miao et al. (2015), where the PPTD framework was presented for enabling privacy-preserving truth discovery in crowdsensing systems. The PPTD framework relies on the threshold Paillier cryptosystem to protect the confidentiality of users' sensory data and weights. In particular, the cloud server in the PPTD framework is responsible for aggregating any encrypted intermediate data values across users, based on the homomorphic

property of the Paillier cryptosystem. And a number of users are required to assist the cloud server in the decryption of any encrypted intermediate aggregate values. Since the PPTD framework heavily relies on the cost-expensive threshold Paillier cryptosystem and requires a number of users for decryption assistance, it does not perform well in performance, incurring considerable computation and communication overhead. Later, some follow-up works (Xu et al. 2017; Zheng et al. 2017b; Miao et al. 2017) were presented, which proposed different security designs so as to achieve better performance than the PPTD framework. The work by Xu et al. (2017) achieved better performance at the price of sharing a secret key among all the users, which can make the system brittle in the face of user compromise. The work by Zheng et al. (2017b) mainly built on a lightweight aggregation protocol which neither requires the sharing of a secret key among users nor needs to rely on an additional party for key setup. The work by Miao et al. (2017) operated under a two-cloud-server model and relied on additively homomorphic encryption for supporting efficient and privacy-preserving truth discovery.

Despite the effectiveness, the designs in all the above works required users to keep staying online for active participation throughout the whole privacy-preserving truth discovery procedure. Considering this requirement may not always be satisfied in practice, Tang et al. (2018) proposed a design enabling noninteractive privacy-preserving truth discovery, where users do not need to stay online to actively participate in the truth discovery procedure after they submit the (encrypted) sensory data. Their design was built under a two-server model and mainly relied on the classical secure two-party computation technique called garbled circuits to realize privacy-preserving truth discovery on the cloud side. And several optimization techniques were presented to customize their design to achieve good efficiency.

In addition to the above works which proposed designs tailored for static data, there are also some other works that focused on different application contexts. For example, Zheng et al. (2018)

targeted crowdsensing applications with long-tail data, where the number of sensing tasks performed by different users may vary significantly. They proposed a design to support encrypted confidence-aware truth discovery as to achieve high accuracy in estimating the ground truth for long-tail data. Cai et al. (2018) targeted crowdsensing applications with data streams, where sensory data are collected in a streaming fashion. Namely, the new data continue to sequentially arrive over time. They proposed a design to support privacy-preserving streaming truth discovery, in which data are scanned only once and no iterative computation over the same data records is required.

Key Applications

Private truth discovery in cloud-empowered crowdsensing systems can be beneficial to mine truthful information in a privacy-preserving manner from applications, where different mobile users report conflicting measurements about the same sensing objects, such as traffic conditions, broken public utilities, gas prices, weather conditions, and air quality.

Cross-References

- [IoT Security](#)
- [Mobile Security and Privacy](#)
- [Wireless Data Security](#)

References

- Cai C, Zheng Y, Wang C (2018) Leveraging crowdsensed data streams to discover and sell knowledge: a secure and efficient realization. In: Proceedings of IEEE international conference on distributed computing systems, Vienna, 2–6 July 2018
- Ganti RK, Ye F, Lei H (2011) Mobile crowdsensing: current state and future challenges. *IEEE Commun Mag* 49(11):32–39
- Li Y, Gao J, Meng C, Li Q, Su L, Zhao B, Fan W, Han J (2015) A survey on truth discovery. *SIGKDD Explor* 17(2):1–16

- Li Y, Li Q, Gao J, Su L, Zhao B, Fan W, Han J (2016) Conflicts to harmony: a framework for resolving conflicts in heterogeneous data by truth discovery. *IEEE Trans Knowl Data Eng* 28(8):1986–1999
- Miao C, Jiang W, Su L, Li Y, Guo S, Qin Z, Xiao H, Gao J, Ren K (2015) Cloud-enabled privacy-preserving truth discovery in crowd sensing systems. In: *Proceedings of ACM conference on embedded networked sensor systems*, Seoul, 1–4 November 2015
- Miao C, Su L, Jiang W, Li Y, Tian M (2017) A lightweight privacy-preserving truth discovery framework for mobile crowd sensing systems. In: *2017 IEEE conference on computer communications, INFOCOM 2017*, Atlanta, 1–4 May 2017, pp 1–9
- Tang X, Wang C, Yuan X, Wang Q (2018) Non-interactive privacy-preserving truth discovery in crowd sensing applications. In: *Proceedings of IEEE conference on computer communications*, Honolulu, 15–19 April 2018
- Xu G, Li H, Tan C, Liu D, Dai Y, Yang K (2017) Achieving efficient and privacy-preserving truth discovery in crowd sensing systems. *Comput Secur* 69: 114–126
- Zheng Y, Cui H, Wang C, Zhou J (2017a) Privacy-preserving image denoising from external cloud databases. *IEEE Trans Inf Forensics Secur* 12(6): 1285–1298
- Zheng Y, Duan H, Yuan X, Wang C (2017b) Privacy-aware and efficient mobile crowdsensing with truth discovery. *IEEE Trans Dependable Secure Comput* (99):1–1. <https://doi.org/10.1109/TDSC.2017.2753.245>
- Zheng Y, Yuan X, Wang X, Jiang J, Wang C, Gui X (2017c) Toward encrypted cloud media center with secure deduplication. *IEEE Trans Multimedia* 19(2):251–265
- Zheng Y, Duan H, Wang C (2018) Learning the truth privately and confidently: encrypted confidence-aware truth discovery in mobile crowdsensing. *IEEE Trans Inf Forensics Secur* 13(10):2475–2489

Proofs of Stake

- [Blockchain](#)

Proofs of Work

- [Blockchain](#)

Prospect Theory

- [Prospect Theory for Communication Networks](#)

Prospect Theory for Communication Networks

Man Hon Cheung and Junlin Yu
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China

Synonyms

[Behavioral economics](#); [Expected utility theory](#); [Prospect theory](#)

Definitions

Prospect theory (PT) is a Nobel-prize winning behavioral economic theory proposed by Daniel Kahneman and Amos Tversky. It explains the decisions among alternatives under uncertainty by taking into account the human psychology. It generalizes the expected utility theory (EUT), which fails to explain a collection of observations involving risky choices.

Historical Background

In a wireless network, an agent (e.g., a mobile user) often needs to face *uncertainty* when allocating the resources. Externally, in the shared wireless spectrum, he may encounter channel fluctuations (e.g., fading and shadowing) and interference from other agents, when he chooses the power level to transmit or the wireless channel to access, respectively. Internally, he may encounter uncertainty of his future data demand, which affects his choice of data plan and the consumption of his data quota.

Traditionally, the *expected utility theory* (EUT) is widely adopted in the literature for wireless resource allocation. In EUT, an agent aims to maximize the weighted average of his utilities under different uncertain outcomes. Mathematically, with uncertainty, there are I possible outcomes, where α_i is the i th outcome and p_i is the corresponding probability that it occurs. Let y be the decision variable on the resource allocation and $\mathcal{Y}$ be its feasible set. Let $f(y, \alpha_i)$ be the *utility* of the decision-maker when his decision is y and the realized outcome is α_i . For example, in wireless communications, $f(y, \alpha_i)$ may represent the agent's throughput when the transmit power is y and the channel condition is α_i . Under EUT, the decision-maker maximizes his expected utility by solving

$$\max_{y \in \mathcal{Y}} \sum_{i=1}^I f(y, \alpha_i) p_i. \quad (1)$$

While the EUT aims to optimize the agent's expected performance, it may not be able to accurately capture the realistic decision-making if the agent is a *human being*, who is influenced by his cognitive limitations and risk preferences. First, a human decision-maker may not always be a *fully rational* agent, who maximizes his expected utility. Instead, he is often *bounded rational* and cannot maximize his expected utility due to his cognitive limitations. Moreover, it was reported in Kahneman and Tversky (1979) that humans are often *risk-averse* when gaining utility but *risk-seeking* when losing utility, where this evaluation is relative to a neutral reference point.

As a result, Daniel Kahneman and Amos Tversky proposed the *prospect theory* (PT) (Kahneman and Tversky 1979) that generalizes EUT to explain these additional observations. Later on, in 2002, Daniel Kahneman was awarded the Nobel Prize in economics "for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty." It has then been widely applied in many areas, which includes finance (Jin and Zhou 2008), consumption decisions (Koszegi

and Rabin 2009), industrial organizations (Paul and Koszegi 2012), labor supplies (Camerer and Loewenstein 2004), and political sciences (Levy 2003).

However, the study of resource allocation in communication networks based on PT only emerged recently. In this article, the mathematical formulation of PT, its key applications in communication networks, and some open research problems will be discussed.

Prospect Theory

In this section, the three key features in the mathematical modeling of PT, namely, reference point, s-shaped value function, and probability distortion (Kahneman and Tversky 1979, 2000), are discussed in details. The cognitive features behind them are highlighted.

Reference Point

The EUT assumes that an agent's level of satisfaction is characterized by the expected utility that he achieves. In other words, if two agents receive the same expected utility, they should be equally happy. However, as humans, our level of satisfaction is also related to the outcome that we expect (Kahneman 2011): an outcome above our expectation is a gain that makes us happy, while an outcome below our expectation is a loss that disappoints us.

Let R_p be the *reference point*, which indicates the agent's expectation of the outcome. The agent considers an outcome a *gain* if it is higher than the reference point and a *loss* if it is lower than the reference point. Thus, having a higher reference point means that the agent is more likely to treat an outcome as a loss, while having a lower reference point means that he is more likely to treat an outcome as a gain.

S-Shaped Asymmetrical Value Function

Besides the impact of the reference point, there are two other important cognitive features at the heart of PT, namely, diminishing sensitivity and loss aversion (Kahneman 2011).

Diminishing sensitivity reflects a fundamental feature of human cognition, where the marginal change in the perception diminishes with respect to the distance from the reference point. For example, regarding the sensory perception, a small increment of light intensity in a dark room is more detectable than that in an illuminated room. Regarding the change of wealth, the difference between a gain of \$200 and a gain of \$300 appears to be larger than that between a gain of \$2200 and a gain of \$2300. Similarly, the difference between a loss of \$200 and a loss of \$300 appears to be larger than that between a loss of \$2200 and a loss of \$2300 (Kahneman and Tversky 1979). In other words, a decision-maker often experiences a diminishing marginal utility when evaluating a gain and a diminishing marginal disutility when evaluating a loss.

Loss aversion refers to the human's stronger preference of avoiding losses than achieving gains. This asymmetry between the power of positive and negative experiences has an evolutionary history, where organisms that treat threats as more urgent than opportunities have a better chance to survive (Kahneman 2011).

Putting it together, due to the diminishing sensitivity, the *value function* $v(x)$, which maps an objective outcome x to the agent's subjective valuation, is concave in the gain region (i.e., when $x \geq 0$) and convex in the loss region (i.e., when $x < 0$) under a zero reference point. Moreover, due to loss aversion, it is asymmetrical around the reference point, where the function is steeper for losses than for gains (i.e., $|v(-x)| > v(x)$ for any $x > 0$). The result is an s-shaped asymmetrical value function $v(x)$ as illustrated in Fig. 1a. A commonly used value function in the PT literature in the form of Tversky and Kahneman (1992) and Kahneman and Tversky (2000)

$$v(x) = \begin{cases} x^\beta, & \text{if } x \geq 0, \\ -\lambda(-x)^\gamma, & \text{if } x < 0, \end{cases} \quad (2)$$

is shown, where $\lambda \geq 1$, $0 < \beta \leq 1$, and $0 < \gamma \leq 1$. Notice that all the outcomes are measured relative to the reference point R_p , which is normalized to $R_p = 0$ in the figure. The

parameter λ is the loss penalty parameter, where a larger λ indicates that the agent is more afraid of a loss. The parameters β and γ are the risk parameters, where the value function of the gain part is more concave (i.e., the decision-maker is more risk-averse) when β approaches zero, and the value function of the loss part is more convex (i.e., the decision-maker is more risk-seeking) when γ approaches zero. It should be noted that EUT is a special case when $\lambda = 1$ and $\gamma = \beta = 1$, where the value function becomes a linear function of $v(x) = x$.

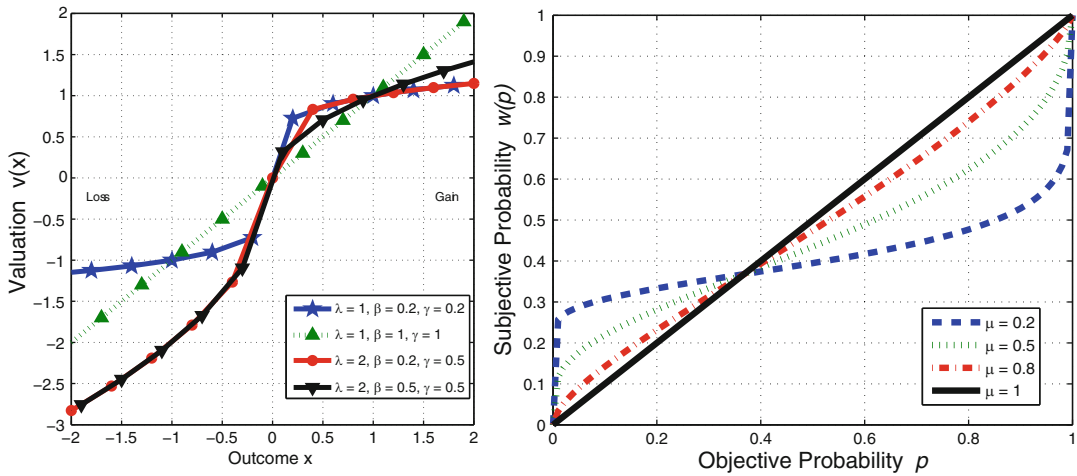
As a result, combining the impact of both the reference point and the s-shaped value function, an agent is more likely to encounter losses under a high reference point R_p and is thus more risk-seeking. On the other hand, under a low reference point R_p , the agent is more likely to encounter gains and is thus more risk-averse.

Probability Distortion

It was shown that the decision weight people assign to an outcome is not identical to the probability of that outcome. For example, consider the following four events of a 5% improvement in winning a big prize: (a) from 0% to 5%, (b) from 5% to 10%, (c) from 60% to 65%, and (d) from 95% to 100%. It was shown that events (a) or (d) are more significant than events (b) or (c) (Kahneman 2011).

First, the larger impact of event (a) is due to the *possibility effect*, which causes low probability events to be weighted more than their objective probability. It can be used to explain the gambling behavior, where people are willing to buy lottery tickets for a very low probability of winning a big prize. Second, the large impact of event (d) is due to the *certainty effect*, which causes events that are almost certain (e.g., 95% chance of winning a prize) to be weighted less than their objective probability. To sum up, human tends to *overweight low probability events* but *underweight high probability events*.

Figure 1b illustrates the probability distortion function $w(p)$, which captures humans' psychological overweighting of low probability events and underweighting of high probability events (Kahneman and Tversky 1979). A commonly



Prospect Theory for Communication Networks, Fig. 1 (a) The s-shaped asymmetrical value function $v(x)$ with reference point $R_p = 0$. (b) The probability distortion function $w(p)$

used probability distortion function is Prelec and Loewenstein (1991)

$$w(p) = \exp(-(-\ln p)^\mu), \quad 0 < \mu \leq 1, \quad (3)$$

where p is the *objective* probability of an outcome and $w(p)$ is the corresponding *subjective* probability. Here, μ is the probability distortion parameter, which reveals how a person's subjective evaluation distorts the objective probability. A *smaller* μ means a *larger* distortion. When $\mu = 1$, it refers to the case of EUT without probability distortion that $w(p) = p$.

Optimization Problem

Combining the three features in PT, the decision-maker aims to solve

$$\max_{y \in \mathcal{Y}} \sum_{i=1}^I v(f(y, \alpha_i) - R_p) w(p_i), \quad (4)$$

which is a *non-convex* optimization problem in general due to the s-shaped value function $v(x)$. In reality, an operator's decision is influenced by not only the consideration of the expected utility maximization but also by the level of his *risk preference*, and they are both captured in problem (4). As a result, problem (4) is more general than problem (1).

Key Applications

In this section, the literature that applied prospect theory to the resource allocation problems involving uncertainties in communication networks and smart grids is reviewed as follows:

- *Wireless random access*: The first paper of this subject is due to Li and Mandayam (2012), who compared the equilibrium strategies of a two-user random access game under EUT and PT. A linear value function was considered. They extended the analysis in Li and Mandayam (2014) to a general random access channel model with more users.
- *Spectrum investment in cognitive radio networks*: Yu et al. (2016) considered the spectrum investment of a virtual operator in sensing and leasing. Although sensing is usually cheaper than leasing, the amount of available spectrum obtained by sensing is *uncertain* due to the primary users' activities in the licensed band. The authors showed that compared to an EUT operator, both the risk-averse and risk-seeking operator achieve a smaller expected profit. On the other hand, a risk-averse operator can guarantee a larger minimum possible profit, while a risk-seeking operator can achieve a larger maximum possible profit.

- *Mobile data trading*: Yu et al. (2017) considered a mobile data trading market, where users can trade their monthly 4G mobile data quota with each other. In this market, a user needs to make the trading decision under the *uncertainty* of his future data usage in the remaining time of the billing cycle. By comparing with the EUT benchmark, the authors showed that a PT user with a lower reference point is more risk-averse and is thus more willing to buy mobile data to reduce the risk that his future data demand will exceed the quota. Moreover, when the probability of having a high future data demand is low, a PT user is more willing to buy mobile data due to the probability distortion comparing with an EUT user.
- *Pricing*: Yang et al. (2015) considered the impact of end-user decision-making on wireless resource pricing, where there is an uncertainty in the quality of service (QoS) guarantees. They designed the data pricing and channel allocation algorithms based on PT.
- *Smart grid*: Wang et al. (2014) formulated a noncooperative game among consumers in an energy exchange system. They applied PT to explicitly account for the users' subjective perceptions of their expected utilities. Xiao et al. (2015) studied the static energy exchange game among microgrids that are connected to a backup power plant. They analyzed the Nash equilibria under various scenarios based on PT and evaluated the impact of user's objective weight on the equilibrium of the game.

Open Problems

In this section, some open problems for the applications of PT to communication networks are outlined.

- *Shifting reference point*: The literature mentioned above mainly considered the *static* user decision in one time slot. It is interesting to consider the *dynamic* decisions in multiple

time slots, where a user's reference point may change over time.

- *Risk preference estimation*: In order to solve problem (4), it is important to know the agent's risk preference accurately. Yu et al. (2017) proposed an algorithm to estimate the parameters in the value function. Nevertheless, further studies are needed to accurately estimate the reference point and probability distortion function as well.

Cross-References

► Preferences and Utility Functions

References

- Camerer CF, Loewenstein G (2004) Advances in behavioral economics. Princeton University Press, Princeton
- Jin H, Zhou X (2008) Behavioral portfolio selection in continuous time. *Math Finance* 18(3):385–426
- Kahneman D (2011) Thinking, fast and slow. Farrar, Strauss and Giroux, New York
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47(2):263–291
- Kahneman D, Tversky A (2000) Choices, values, and frames. Cambridge University Press, New York
- Koszegi B, Rabin M (2009) Reference-dependent consumption plans. *Am Econ Rev* 99(3):909–936
- Levy JS (2003) Applications of prospect theory to political science. *Synthese* 135(2):215–241
- Li T, Mandayam NB (2012) Prospects in a wireless random access game. In: Proceedings of CISS, Princeton
- Li T, Mandayam NB (2014) When users interfere with protocols: prospect theory in wireless networks using random access and data pricing as an example. *IEEE Trans Wireless Commun* 13(4):1888–1907
- Paul H, Koszegi B (2012) Regular prices and sales. <http://emlab.berkeley.edu/~botond/rps.pdf>
- Prelec D, Loewenstein G (1991) Decision making over time and under uncertainty: a common approach. *Manag Sci* 37(7):770–786
- Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertainty* 5(4):297–323
- Wang Y, Saad W, Mandayam NB, Poor HV (2014) Integrating energy storage into the smart grid: a prospect theoretic approach. In: Proceedings of IEEE ICASSP, Florence
- Xiao L, Mandayam N, Poor HV (2015) Prospect theoretic analysis of energy exchange among microgrids. *IEEE Trans Smart Grid* 6(1):63–72

- Yang Y, Park L, Mandayam N, Seskar I, Glass A, Sinha N (2015) Prospect pricing in cognitive radio networks. *IEEE Trans Cogn Commun Netw* 1(1):56–70
- Yu J, Cheung MH, Huang J (2016) Spectrum investment under uncertainty: a behavioral economics perspective. *IEEE J Sel Areas Commun* 34(10):2667–2677
- Yu J, Cheung MH, Huang J, Poor HV (2017) Mobile data trading: behavioral economics analysis and algorithm design. *IEEE J Sel Areas Commun* 35(4):994–1005

Protocol Stack Architecture for Wireless Sensor Networks

Quanxin Zhao

Department of Technology Center, Sichuan Netop Telecom Co., Ltd., Mianyang, Sichuan, China

Synonyms

[Network stack for wireless sensor networks](#); [Protocol stack for wireless sensor networks](#)

Definition

Wireless sensor network:

Wireless sensor network (WSN) refers to a group of spatially dispersed and dedicated sensors for monitoring and recording the physical conditions of the environment and organizing the collected data at a central location. WSNs measure environmental conditions like temperature, sound, pollution levels, humidity, wind, and so on.

OSI model:

The Open Systems Interconnection model (OSI model) is a conceptual model that characterizes and standardizes the communication functions of a telecommunication or computing system without regard to its underlying internal structure and technology. Its goal is the interoperability of diverse communication systems with standard protocols. The model partitions a communication system into abstraction layers.

The original version of the model defined seven layers.

Protocol stack:

The protocol stack or network stack is an implementation of a computer networking protocol suite or protocol family. The terms are often used interchangeably; strictly speaking, the suite is the definition of the Communications protocols, and the stack is the software implementation of them.

Historical Background

WSNs are designed for gathering information in a large geographical area. Massive sensors are networked together as a wireless sensor network to strengthen the power of sensing. Such typically self-configuring network is often controlled through a software core engine. To demonstrate the important engine structure of WSNs, protocol stack architecture is present.

ISO, The International Organization for Standardization, promotes worldwide standards for commercial, industrial, and proprietary, acting as an international standard-setting body. A standard is an agreed or required level of quality or attainment. ISM (The Industrial, Scientific and Medical) radio bands established by ITU (the International Telecommunications Union) finds its applications into microwave ovens, medical diathermy machines, radio-frequency process heating, etc. Taking both the low power consumption of WSNs and the ISM frequency bands into consideration, WSN standards from ISO have developed the functions and protocols to interface with other wired or wireless networks.

The IEEE 802.15 working group has many different approved standards (represented by 802.15.x) related to WSNs. Currently, all 802.15.x approved standards have proposed only physical (PHY) and medium access control (MAC) layers, leaving the design of network, transport, and application layers for other parties.

At present, WSN standards (Hossam 2016) include: ANT, EnOcean, IEEE 802.15.4 Low Rate WPANs, IEEE 802.15.3, Impulse

Radio Ultra-Wide Bandwidth Technology 802.15.4a, INSTEON, ISA100.11a, MyriaNed, Wibree(BLE), Wavenis, WirelessHART, ZigBee, Z-Wave, 6LoWPAN, etc.

Key Applications

WSNs have many key applications, which can be categorized into daily life, environmental, health, industrial, military, multimedia, and surveillance. Specifically speaking, WSNs can be used for biological or chemical detection, health care monitoring, public security, system control, target tracking, traffic control, etc.

As WSNs have spanned a wide range of important applications, it is necessary to satisfy the diverse requirements of applications by considering some design goals. These goals are application-specific design, adaptive network operation, data fusion and localized processing, fault tolerance and reliability, low-cost and small sensor nodes, resource-efficient design, scalable architectures and efficient protocols, self-configuration and self-organization, secure design, and time synchronization.

Differences with Other Networks

Similar to a general wireless network, a WSN also transmits a packet with a header among sensor nodes. But the existing algorithms designed for general wireless networks, such as ad hoc networks, cannot be directly applied for WSNs (Ian et al. 2006; Jennifer et al. 2008), which is due to some reasons:

1. Broadcast transmission is used due to restricted memory space and limited power supply, compared with point-to-point communication in general wireless networks.
2. Sensor nodes are with both the larger numbers and the higher probability of permanent failure, compared with a typical ad-hoc network.

Protocol Stack Architecture

WSN protocol stack design is based on the OSI seven-layer model. This model is designed by the International Organization for Standardization (ISO) and consists of physical layer, data link layer, network layer, transport layer, session layer, presentation layer, and application layer.

Since too many layers of OSI model can bring WSNs overlay complex and implement difficulty, WSNs cannot take all the design requirement of the OSI model. Therefore, the WSN protocol stack has adopted only five layers. The protocol stack architecture of WSNs (Shuang-Hua 2014) in Fig. 1 consists of the physical layer, data link layer, network layer, transport layer, and application layer, backed by a power management plane, mobility management plane, task management plane, quality-of-service (QoS) management plane, and security management plane.

The WSN protocol stack consists of five layers. Each layer has a specific set of task and works independently.

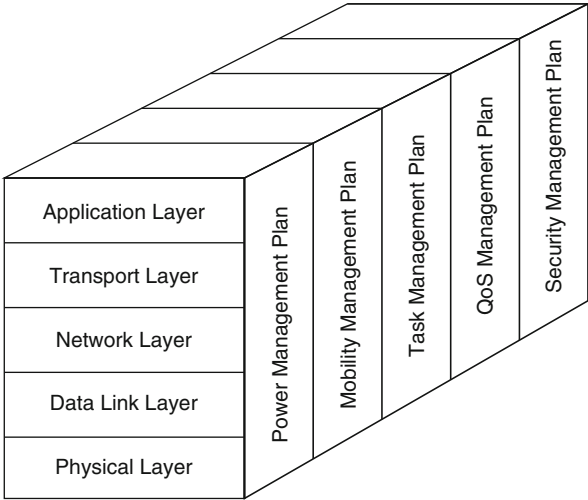
The responsibility of the first layer, namely, the physical layer, includes

1. Defining the cable-compatible connector for sensor nodes.
2. Managing the connections among sensor nodes and their communication transmission medium.
3. Managing carrier frequency, data encryption, frequency selection, robust modulation, signal detection, signal transmission, and break reception.

The responsibility of the second layer, namely, the data link layer, includes

1. Providing services such as error detection and correction, MAC, reliable delivery.
2. Enabling massive nodes successfully accessing and fully sharing communication transmission medium.
3. Minimizing packet collision with neighboring sensors.

**Protocol Stack
Architecture for Wireless
Sensor Networks, Fig. 1**
Protocol stack architecture
of WSN



The responsibility of the third layer, namely, the network layer, includes

- 1. Establishing communication transmission paths between sensor nodes and routing packets along the established paths successfully.
- 2. Designing routing protocols with different requirements for diverse paths setup policy, such as the best QoS paths, the best lifetime paths.

The responsibility of the fourth layer, namely, the transport layer, includes

- 1. Providing with transparent and reliable communications between end sensors, i.e., the transmitter and receiver.
- 2. Considering transport layer protocol forms, such as the transmission control protocol (TCP) and the user datagram protocol (UDP).
- 3. Providing a reliable communication service with connection-oriented transport layer protocols, such as TCP, for the aim of extensive error handling, flow control, transmission control.
- 4. Providing an unreliable communication service with connectionless transport layer protocols, such as UDP.

The responsibility of the fifth layer, namely, the application layer, includes

- 1. Designing applications including, File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP), Simple Network Management Protocol (SNMP), Secure Shell (SSH), and Telnet.
- 2. Dealing with sensed information, data encryption, data formatting, and storage.
- 3. Scanning and detecting the availability of network resources for the underlying layers.

The WSN protocol stack is backed by five management function planes for the integration of the administration, configuration, maintenance, and operation.

The responsibility of the power management plane includes

- 1. Monitoring the sensor node’s current power level and the amount of power consumption.
- 2. Turning off functionality to preserve energy when the power level is below some threshold.

The responsibility of the mobility management plane includes

- 1. Handling WSN layout issues.
- 2. Managing sensor node location and network activity distribution.
- 3. Detecting movement of nodes to maintain node to node communication and data route.

The responsibility of the task management plane includes

1. Assigning sensing tasks to necessary nodes in the sensing field.
2. Scheduling sensor nodes without sensing tasks for the optimization of data routing and aggregation.

The responsibility of the QoS management plane includes

1. Managing data latency, data acquisition accuracy, fault tolerance, system performance.
2. Determining QoS metrics to deal with performance optimization, error control, and fault tolerance.

The responsibility of the security management plane includes

1. Handling functionalities such as authentication, encryption, threat detection and recovery, intrusion detection.
2. Controlling sensor nodes to access sensitive data.

Cross-References

- [Wireless Sensor Networks](#)

References

- Hossam MAF (2016) Wireless sensor networks: concepts, applications, experimentation and analysis. Springer, Singapore
- Ian FA, Tommaso M, Kaushik RC (2006) A survey on wireless multimedia sensor networks. Comput Netw. <https://doi.org/10.1016/j.comnet.2006.10.002>
- Jennifer Y, Biswanath M, Dipak G (2008) Wireless sensor network survey. Comput Netw 52:2292–2330
- Shuang-Hua Y (2014) Wireless sensor networks: principles, design and applications. Springer, London

Protocol Stack for Wireless Sensor Networks

- [Protocol Stack Architecture for Wireless Sensor Networks](#)

Provide Synonyms to the Article Title

- [Bluetooth Low Energy \(BLE\) Security and Privacy](#)

Providing QoS Support at the Distributed Wireless MAC Protocol

- [QoS-Aware MAC](#)

Proxy Mobile IPv6

Sri Gundavelli
Cisco Systems, San José, CA, USA

Synonyms

[Binding](#); [IP mobility](#); [Local mobility anchor](#); [Mobile access gateway](#); [Mobile node](#); [Mobility session](#); [Proxy mobile IPv6](#)

Background

IP mobility protocols are designed to allow mobile nodes to remain reachable while moving around in the Internet. There are two types of

mobility management approaches: client-based and network-based.

Client-Based Mobility Management – The mobile node is required to be involved in the mobility management. Exchange of IP mobility signaling messages between the mobile node and a mobility anchor in the network enables the creation and maintenance of a binding between the mobile node's home address and its care-of address. It essentially allows the mobility anchor node to tunnel the traffic bound to the mobile node's home address to its care-of address. IP mobility support for IPv6 hosts is specified in Mobile IPv6 specification [RFC-3775].

Network-Based Mobility Management – There is another approach to solving the IP mobility problem and is the network-based approach. This mobility management approach does not require the mobile node to be involved in the exchange of signaling messages between itself and a mobility anchor located in the network. A proxy mobility agent in the network tracks the mobile node's movements and performs the signaling with the mobility anchor on behalf of the mobile node. This essentially removes the requirement of client involvement in mobility management. Proxy Mobile IPv6 (PMIPv6) is a network-based mobility management protocol standardized by IETF. The base protocol is specified in RFC 5213 and RFC 5844; other protocol extensions to Proxy Mobile IPv6 are specified in additional RFC specifications.

Protocol Overview

Proxy Mobile IPv6 is a protocol for building a common and access-technology agnostic mobility core. The key capability of Proxy Mobile IPv6 is its ability to provide mobility management support to a mobile node without requiring its participation in any IP mobility-related signaling. The mobility entities in the network will track the mobile node's movements and will initiate the mobility signaling and set up the required routing state in the network.

The core functional entities in the Proxy Mobile IPv6 domain are the local mobility anchor (LMA) and the mobile access gateway (MAG).

The local mobility anchor is responsible for maintaining the mobile node's reachability state and it is the topological anchor point for the mobile node's home network. All the mobile node's inbound and outbound IP traffic is always routed through the local mobility anchor.

The mobile access gateway is the entity that performs the mobility management on behalf of the mobile node, and it resides on the access link where the mobile node is anchored. The mobile access gateway is responsible for detecting the mobile node's movements to and from the access link and for initiating binding registrations to the mobile node's local mobility anchor.

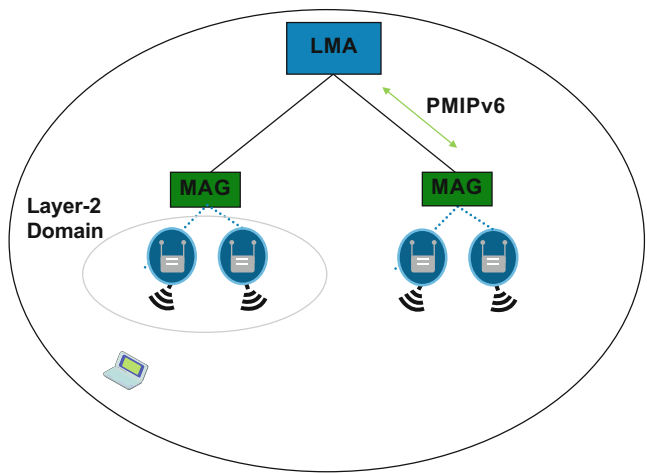
The following diagram is a representation of a mobility management domain based on Proxy Mobile IPv6 (Fig. 1).

Proxy Mobile IPv6 is a control-plane-driven protocol, which uses the control-plane signaling to establish dynamic tunnels between the MAG and LMA. The mobility entities use control plane for session management and for user-plane management. Mobility messages such as PBU/PBA, BRI/BRA, HB, UPN/UPA are used for triggering the protocol peers for activating various functions on the mobile node's binding state. The user plane is based on standard layer-3 encapsulation and routing mechanisms and is used for forwarding the mobile node's IP packets. The supported tunnel encapsulation types are IP-GRE, IP-in-IP, and IP-UDP. The encapsulation mode for the tunnel is negotiated as part of the Proxy Mobile IPv6 signaling messages.

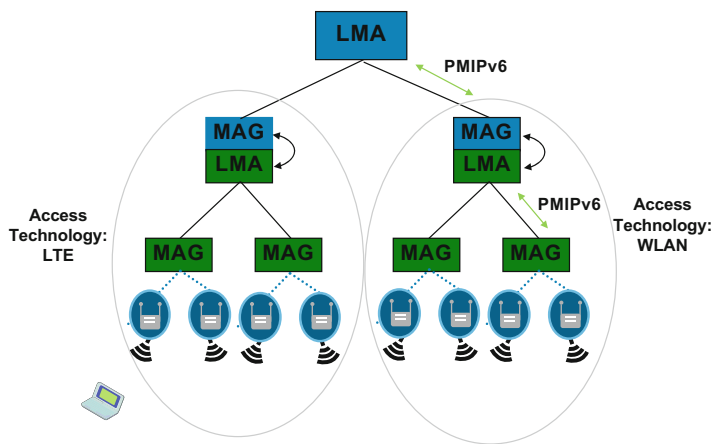
The MAG tracks the mobile node's movements (Attach/Detach triggers) on the access link based on ARP, DHCP, IPv6 Neighbor Discovery, and link-layer specific protocols (e.g., NAS signaling in LTE). Furthermore, the mobile node obtains the IP address configuration from the MAG using DHCP and/or IPv6 Neighbor Discovery.

Proxy Mobile IPv6 protocol is used for network-based mobility management in various

Proxy Mobile IPv6, Fig. 1
Proxy Mobile IPv6 domain



Proxy Mobile IPv6, Fig. 2
Mobility chaining



mobile architectures including 3GPP, 3GPP2, and WiMAX. The protocol is specified over various network interfaces in their system architectures. The protocol is also used for supporting various IP mobility use-cases in enterprise, transportation, and other vertical markets. Applications such as mobile networks, mobile local loops, Wi-Fi Traffic Offload, Connected Car management, and Community Wi-Fi networks are some of the examples where this protocol is deployed.

Mobility Chaining

Mobility domains based on Proxy Mobile IPv6 can be chained in a hierarchical fashion. Chaining allows localization of the mobility events. The mobile node’s session is anchored on the top-level LMA and also on the local LMA in the access network. The mobile node’s movements

within the access network are hidden from the top-level LMA. The user-plane traffic for the mobile node is always routed through the LMA in the top level.

Mobility domain chaining can be achieved by enabling LMA and MAG functions on the same node with some interworking between the two functions for session correlation. Chaining is typically suited for large scale mobility deployments connecting mobility domains based on various access technologies under a common mobility core (Fig. 2).

PMIPv6 Protocol Operation

When a mobile node enters a Proxy Mobile IPv6 domain and attaches to an access link, the mobile access gateway on that access link, after identi-

fyng the mobile node and acquiring its identity, will determine if the mobile node is authorized for the network-based mobility management service.

If the network determines that the mobile node is authorized for network-based mobility service, the network will ensure that the mobile node using any of the address configuration mechanisms permitted by the network will be able to obtain the address configuration on the connected interface and move anywhere in that Proxy Mobile IPv6 domain. The obtained address configuration includes the address(es) from its home network prefix(es), the default-router address on the link, and other related configuration parameters. From the perspective of each mobile node, the entire Proxy Mobile IPv6 domain appears as a single link, the network ensures that the mobile node does not detect any change with respect to its layer-3 attachment even after changing its point of attachment in the network.

The mobile node may be an IPv4-only node, IPv6-only node, or a dual-stack (IPv4/v6) node. Based on the policy profile information that indicates the type of address or prefixes to be assigned for the mobile node in the network, the mobile node will be able to obtain an IPv4, IPv6, or dual IPv4/IPv6 address and move anywhere in that Proxy Mobile IPv6 domain.

Example Call Flow

The following is an example call flow for a mobile node's attachment to a Proxy Mobile IPv6 network (Fig. 3).

Signaling Plane

The operation of Proxy Mobile IPv6 protocol is primarily based on the protocol signaling between the mobile access gateway and the local mobility anchor. The base protocol specified in RFC 5213, RFC 5844, and other specifications defines the structure and semantics of the interface between the mobile access gateway and the local mobility anchor. Although the protocol does not define any signaling interface

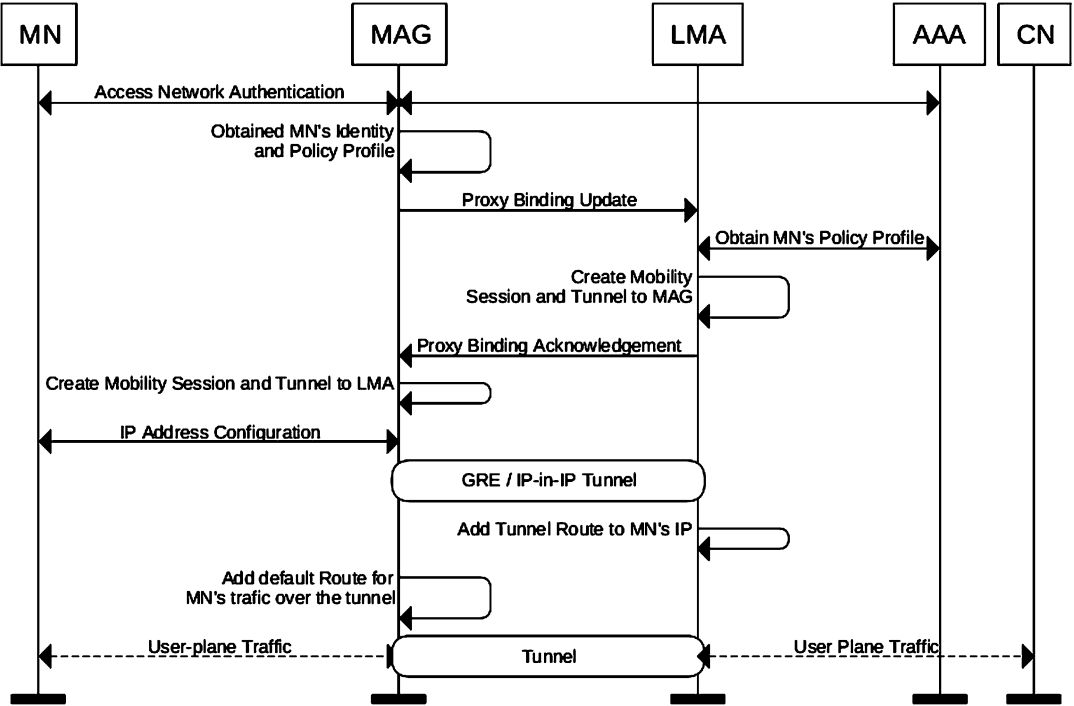
between the mobile node and the mobile access gateway, it is an essential interface needed for the protocol operation. The mobile access gateway relies on standard IP interfaces (e.g., DHCP, IPv6 ND, ARP) and access-technologic specific layer-2 interfaces for learning the mobile node's Attach/Detach events, for learning the mobile node's authenticated identity and for delivering IPv4/IPv6 address configuration to the mobile node.

Protocols Used Between MN to MAG

Protocols	Description	Reference
Layer-2	Link-layer protocols are used for learning the mobile node's attach/detach events on the access link, for learning the mobile node's authenticated identity and for obtaining the mobile node's policy profile	Refer to access specific interfaces (e.g., WLAN/3GPP specifications)
ARP	Used for detecting the mobile node's presence on the access link	RFC-826
IPv6 ND	Used for delivering the IPv6 address configuration	RFC-4861RFC-4862
DHCP	Used for delivering the IPv4/IPv6 address configuration	RFC-2131RFC-3315

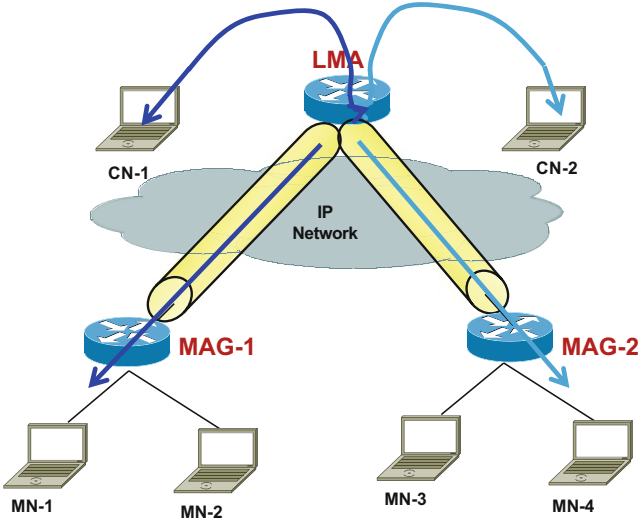
Signaling Messages Between MAG and LMA

Protocol messages	Description	Reference
PBU/PBA	Messages for session management	RFC-5213
BRI/BRA	Messages for binding revocation management	RFC-5846
UPN/UPA	Messages for notifications related to session state	RFC-7077
Heartbeat	Messages for peer state detection	RFC-5847



Proxy Mobile IPv6, Fig. 3 Call flow for PMIPv6 signaling

Proxy Mobile IPv6, Fig. 4
Packet forwarding path



User Plane

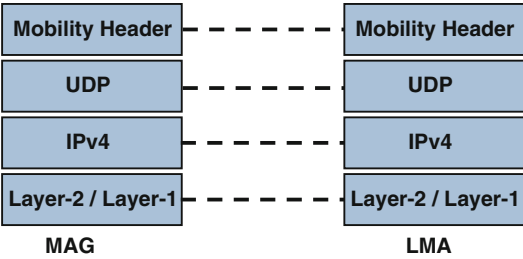
See Fig. 4

Traffic from CN to MN

The mobile node's inbound IP traffic from the correspondent node is always routed through

the local mobility anchor. The IP address that is assigned on the mobile node is topologically anchored on the LMA and so the local mobility anchor is always in the path for the mobile node's traffic.

When a correspondent node sends a packet to the mobile node's home address, the packet



Proxy Mobile IPv6, Fig. 5 IPv4 control-plane stack

hits the local mobility anchor as it advertises the reachability for that prefix associated with that IP address.

The local mobility anchor on receiving the packet will forward the packet through a bidirectional tunnel to the mobile access gateway and which in turn will deliver it to the mobile node on the access link. There will be a prefix route for the mobile node through the bidirectional tunnel.

Traffic from MN to CN

The mobile node’s outbound IP traffic is always tunneled from the mobile access gateway to the local mobility anchor.

When a mobile node sends a packet to a correspondent node, the packet is first received by the first-hop router and which is the mobile access gateway. The mobile access gateway tunnels the packet to the mobile node’s local mobility anchor.

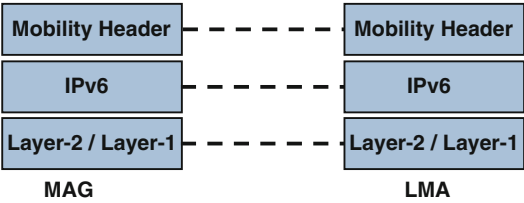
The local mobility anchor on receiving the packet from the tunnel removes the tunnel header and forwards the packet to the correspondent node.

Protocol Stacks

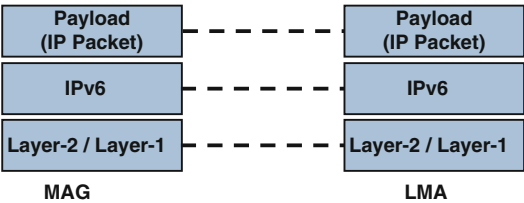
Control-Plane Traffic

The following are the supported control-plane protocol stacks on LMA and MAG.

When using IPv6 transport, the following protocol stack gets activated on the MAG and on the LMA. The Mobility Header is an extension header in the IPv6 packet. However, when IPsec is used between MAG and LMA for securing the



Proxy Mobile IPv6, Fig. 6 IPv6 control-plane stack



Proxy Mobile IPv6, Fig. 7 IPv6 user-plane protocol stack

messages, IPsec ESP header is added right above the IPv6 header (Fig. 5).

When using IPv4 transport, the following protocol stack gets activated on the MAG and the LMA. However, when IPsec is used between MAG and LMA for securing the messages, IPsec ESP header is added right above the IPv4 header and below the UDP header (Fig. 6).

User-Plane Traffic

The following are the supported user-plane protocol stacks on the MAG and the LMA.

When using IP-in-IP encapsulation, the IP packet from the mobile node is encapsulated in an IPv6 packet. However, when IPsec is used between MAG and LMA for securing the user-plane traffic, IPsec ESP header is added right above the IPv6 header (Fig. 7).

When using GRE encapsulation, the IP packet from the mobile node is encapsulated in a GRE packet. GRE is supported both on IPv4 and IPv6 transports. However, when IPsec is used between MAG and LMA for securing the user-plane traffic, IPsec ESP header is added right above the IPv4 header and below the GRE header (Fig. 8).

Applications

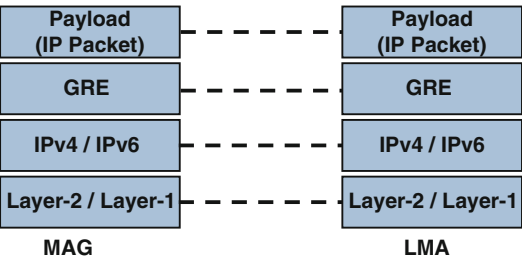
Hybrid Access Support

Cable operators are exploring options to offer higher throughput access to the Internet. This is especially true with the incumbent xDSL service providers with a large deployment of both xDSL and LTE networks. These operators have no immediate opportunity for the expensive FTTH network upgrade or upgrade to cable Internet, but they have to compete with the cable operators who are able to offer much higher throughputs with their DOCSIS 3.x-based access architectures. Customers on legacy networks

have issues viewing high-definition videos. To address this requirement, operators are using Proxy Mobile IPv6 technology with multipath extensions for LTE and DSL link aggregation.

In this deployment, the MAG functionality is collocated on the CPE, and the LMA functionality is collocated on a network node such as BNG.

The approach allows the operator to specify SLA (Latency, Jitter, Loss) on an application basis and with path preferences (LTE preferred, otherwise DSL). The forwarding functions in the anchor and the CPE will constantly measure the path preferences and ensure the application flows are routed on the paths that meet the application SLA (Fig. 9).



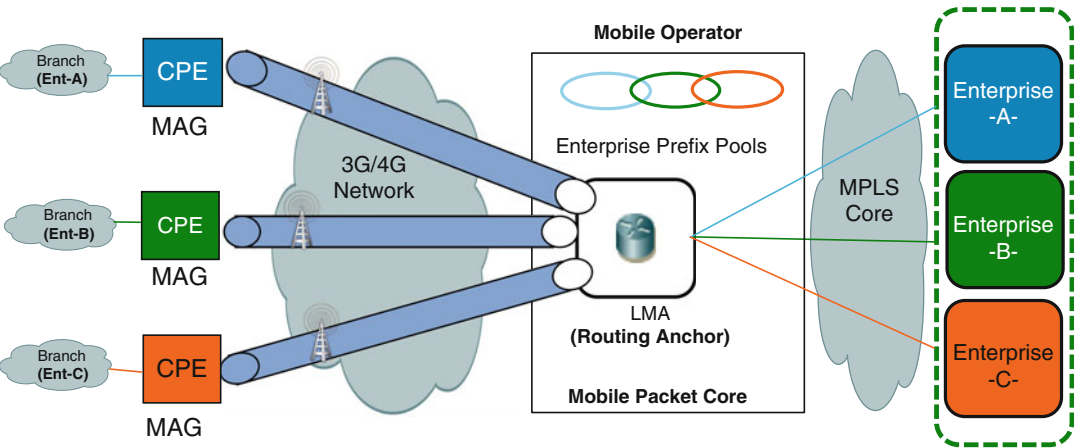
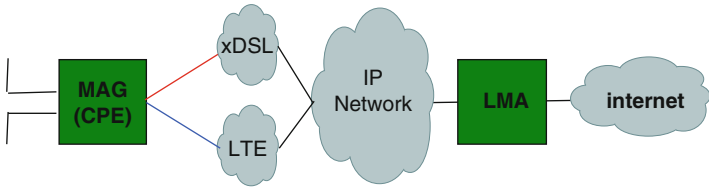
Proxy Mobile IPv6, Fig. 8 IPv4 user-plane protocol stack

Mobile Local Loop

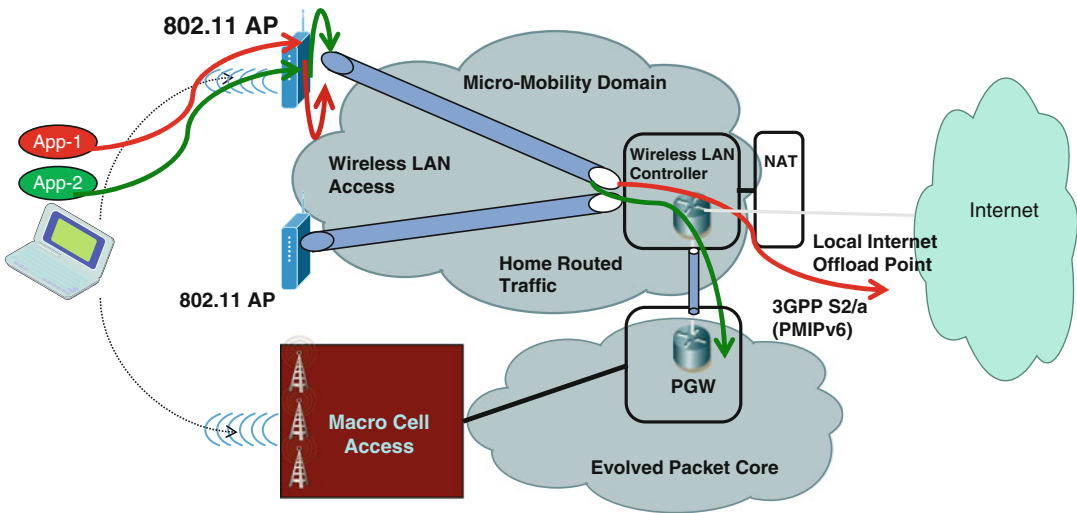
Mobile operators are offering enterprise private routing services over 3G/4G infrastructure. This service allows an enterprise to securely link all their remote small branch offices over a mobile network, without the need for dedicated leased lines, or IPsec/SSL-based VPN's which come with all the provisioning complexity. By leveraging the wireless spectrum, mobile operators are

Proxy Mobile IPv6, Fig. 9

Hybrid access support with Proxy Mobile IPv6



Proxy Mobile IPv6, Fig. 10 Mobile local loop support with Proxy Mobile IPv6



Proxy Mobile IPv6, Fig. 11 WLAN-EPC interworking based on 3GPP S2a Interface

able to offer this private routing service over a 3G/4G network and are able to eliminate any costs associated with the last mile circuit such as payments to local carriers with dedicated leased lines. CPE devices equipped with 3G/4G wireless modules are being used for IP connectivity at the branch office. This new concept is referred to as “Mobile Local Loop.”

Proxy Mobile IPv6 is the key protocol used for enabling this use-case. The MAG function resides in the CPE device and the LMA function with the business logic is enabled on the router in the service provider network (Fig. 10).

WLAN-EPC Interworking

Mobile operators are expanding their network coverage by integrating various access technology domains (e.g., wireless LAN, CDMA, and Long-Term Evolution (LTE)) into a common IP mobility core.

The idea is to provide a consistent set of services and IP mobility support to a mobile node irrespective of the access technology used for the IP connectivity.

For supporting this interworking, the 3GPP has defined interface points allowing integration of Trusted WLAN network with mobile packet core. The 3GPP S2a Proxy Mobile IPv6 [TS23402] reference point, specified by the

3GPP system architecture, defines the protocol interworking for building such integrated multiaccess networks. In this scenario, the mobile node’s IP traffic is always tunneled back from the mobile access gateway in the access network to the local mobility anchor in the home network (Fig. 11).

Terminology

Term	Description
AAA	Authentication, Authorization and Accounting
AP	Access Point
BCE	Binding Cache Entry
BRA	Binding Revocation Acknowledgement
BRI	Binding Revocation Indication
BULE	Binding Update List Entry
GRE	Generic Routing Encapsulation
IP-in-IP	IP Packet Encapsulation in IP Header
LMA	Local Mobility Anchor
MAG	Mobile Access Gateway
MN	Mobile Node
CN	Correspondent Node
NAI	Network Access Identifier
PMIPv6	Proxy Mobile IPv6
RA	Router Advertisement
RS	Router Solicitation

References

- Abinader F, Gundavelli S, Leung K, Krishnan S, Premec D (2012) Bulk registration support in Proxy Mobile IPv6, RFC 6602. <http://www.ietf.org/rfc/rfc6602.txt?number=6602>
- Bernardos CJ et al (2016) Proxy Mobile IPv6 extensions to support flow mobility, RFC 7864. <http://www.ietf.org/rfc/rfc7864.txt?number=7864>
- Devarapalli V, Patel A, Leung K (2007) Mobile IPv6 vendor specific option, RFC 5094. <http://www.ietf.org/rfc/rfc5094.txt?number=5094>
- Devarapalli V, Koodli R, Lim H, Kant N, Krishnan S, Laganier J (2010) Heartbeat mechanism for Proxy Mobile IPv6, RFC 5847. <http://www.ietf.org/rfc/rfc5847.txt?number=5847>
- Gundavelli S (2012) Reserved IPv6 interface identifier for Proxy Mobile IPv6, RFC 6543. <http://www.ietf.org/rfc/rfc6543.txt?number=6543>
- Gundavelli S, Leung K, Devarapalli V, Chowdhury K, Patil B (2008) Proxy Mobile IPv6, RFC 5213. <http://www.ietf.org/rfc/rfc5213.txt?number=5213>
- Gundavelli S, Townsley M, Troan O, Dec W (2011) Address mapping of IPv6 multicast packets on Ethernet, RFC 6085. <http://www.ietf.org/rfc/rfc6085.txt?number=6085>
- Gundavelli S, Korhonen J, Grayson M, Leung K, Pazhyannur R (2012) Access network information option for PMIPv6, RFC 6757. <http://www.ietf.org/rfc/rfc6757.txt?number=6757>
- Gundavelli S, Zhou X, Korhonen J, Koodli R, Feige G (2013) IPv4 traffic offload option for Proxy Mobile IPv6 (SIPTO), RFC 6909. <http://www.ietf.org/rfc/rfc6909.txt?number=6909>
- Keeni G, Koide K, Gundavelli S, Wakikawa R (2012) Proxy Mobile IPv6 management information base, RFC 6475. <http://www.ietf.org/rfc/rfc6475.txt?number=6475>
- Korhonen J, Devarapalli V (2011) Local Mobility Anchor (LMA) discovery for Proxy Mobile IPv6, RFC 6097. <http://www.ietf.org/rfc/rfc6097.txt?number=6097>
- Korhonen J, Gundavelli S, Yokota H, Cui X (2011a) Runtime LMA assignment support, RFC 6463. <http://www.ietf.org/rfc/rfc6463.txt?number=6463>
- Korhonen J, Bournelle J, Chowdhury K, Muhanna A, Meyer U (2011b) MAG & LMA interactions with diameter server, RFC 5779. <http://www.ietf.org/rfc/rfc5779.txt?number=5779>
- Krishnan S, Koodli R, Loureiro P, Wu Q, Dutta A (2012) Localized routing for Proxy Mobile IPv6, RFC 6705. <http://www.ietf.org/rfc/rfc6705.txt?number=6705>
- Krishnan S, Gundavelli S, Liebsch M, Yokota H, Korhonen J (2013) Update notifications for Proxy Mobile IPv6, RFC 7077. <http://www.ietf.org/rfc/rfc7077.txt?number=7077>
- Liebsch M, Seite P, Yokota H, Korhonen J, Gundavelli S (2014) QoS support for Proxy Mobile IPv6, RFC 7222. <http://www.ietf.org/rfc/rfc7222.txt?number=7222>
- Melia T, Gundavelli S (2016) Logical-interface support for IP hosts with multi-access support, RFC 7847. <http://www.ietf.org/rfc/rfc7847.txt?number=7847>
- Muhanna A, Khalil M, Gundavelli S, Leung K (2010a) Generic Routing Encapsulation (GRE) key option for Proxy Mobile IPv6, RFC 5845. <http://www.ietf.org/rfc/rfc5845.txt?number=5845>
- Muhanna A, Khalil M, Gundavelli S, Leung K (2010b) Binding revocation for IPv6 Mobility, RFC 5846. <http://www.ietf.org/rfc/rfc5846.txt?number=5846>
- Patki D, Gundavelli S, Lee J, Fu Q, Bertz L (2017) Mobile access gateway configuration parameters controlled by the local mobility anchor, RFC 8127. <http://www.ietf.org/rfc/rfc8127.txt?number=8127>
- Pazhyannur R, Speicher S, Gundavelli S, Korhonen J, Kaippallimalil J (2015) Extensions to the Proxy Mobile IPv6 (PMIPv6) ANI option, RFC 7563. <http://www.ietf.org/rfc/rfc7563.txt?number=7563>
- Schmidt T, Waehlich M, Krishna S (2011) Base deployment for multicast listener support in Proxy Mobile IPv6, RFC 6224. <http://www.ietf.org/rfc/rfc6224.txt?number=6224>
- Seite P, Yegin A, Gundavelli S (2017) MAG multipath binding option, RFC 8278. <http://www.ietf.org/rfc/rfc8278.txt?number=8278>
- Wakikawa R, Gundavelli S (2010) IPv4 support for Proxy Mobile IPv6, RFC 5844. <http://www.ietf.org/rfc/rfc5844.txt?number=5844>
- Wakikawa R, Pazhyannur R, Gundavelli S, Perkins C (2014) Separation of control and user plane for Proxy Mobile IPv6, RFC 7389. <http://www.ietf.org/rfc/rfc7389.txt?number=7389>
- Xia F, Sarikaya B, Korhonen J, Gundavelli S, Damic D (2012) RADIUS support for Proxy Mobile IPv6, RFC 6572. <http://www.ietf.org/rfc/rfc6572.txt?number=6572>
- Yan Z, Lee J, Lee X (2017) Home network prefix renumbering in Proxy Mobile IPv6, RFC 8191. <http://www.ietf.org/rfc/rfc8191.txt?number=8191>
- Zhou X, Korhonen J, Williams C, Gundavelli S (2014) Prefix delegation for PMIPv6, RFC 7148. <http://www.ietf.org/rfc/rfc7148.txt?number=7148>

Q

QoE Assessment and Measurement

- [QoE Measurement and Assessment of Video Streaming](#)

QoE Assessment Models

Wei Song
University of New Brunswick, Fredericton, NB,
Canada

Synonyms

[QoE metrics](#); [QoE models](#)

Definition

As defined in ITU-T (2008) by the Telecommunication Standardization Sector of International Telecommunication Union (ITU-T), quality of experience (QoE) is “the overall acceptability of an application or service, as perceived subjectively by the end-user.” According to Brunnström et al. (2013) by Qualinet, QoE is defined as “the degree of delight or annoyance of the user of an application or service. It results from the

fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user’s personality and current state.” Correspondingly, QoE assessment models assimilate various network-related performance metrics and application-specific quality metrics to gauge user’s satisfaction toward an application or service.

Historical Background

With the fast development of wireless networks in recent years, multimedia services have become one of the most popular and dominant applications on mobile devices. In particular, mobile video traffic is expected to increase by ninefolds between 2016 and 2021, accounting for over 78% of global data traffic in mobile networks by 2021 (Cisco 2017). Due to the attributes of wireless networks, the quality of multimedia delivery cannot be as good or stable as in a wired network. Traditionally, objective quality of service (QoS) metrics, such as bit error probability (BEP), packet delay, and packet loss, are often used to evaluate multimedia quality. However, QoS is insufficient to gauge user-perceived service quality.

To assess users’ overall satisfaction and acceptance of services, quality of experience (QoE) becomes a stronger indicator than the traditional QoS. As defined in ITU-T (2008) by ITU-T, QoE is “the overall acceptability of an appli-

cation or service, as perceived subjectively by the end-user.” QoE incorporates a wide variety of factors from different layers. Thereby, QoE provides a more comprehensive assessment of user-perceived quality.

Foundations

Based on the classification in Wang et al. (2017), QoE assessment can be categorized into subjective assessment, objective assessment, and hybrid assessment.

Subjective assessment of QoE relies on human users’ direct input on their evaluation for the perceived service quality. Subjective assessment poses strict requirements on experiment design and participant selection. The assessment procedure can be costly and even impractical in some cases. Objective assessment of QoE uses mathematical models to derive the value of QoE from some objectively measurable factors as input variables. The objective assessment methods are more viable although the obtained QoE value is only approximate and not verified by real measurement. Hybrid assessment attempts to employ subjective assessment in a lab environment to calibrate the objective QoE model so that more accurate objective assessment is applied in a larger scale.

Considering the dominance of video traffic in current networks, we take video services as an example to introduce different QoE assessment methods. Traditional IP-based video streaming is usually based on real-time transport protocol (RTP), which runs on top of the user datagram protocol (UDP). The real-time streaming protocol (RTSP) and session description protocol (SDP) are often used together with RTP for session setup and control. Due to issues with firewall and network address translation (NAT) traversals, most videos nowadays are delivered using the hypertext transfer protocol (HTTP) running over the transport control protocol (TCP). The video delivery over HTTP is actually simple bulk download of a video file to the end user’s device. The user experience with progressive downloading is similar to streaming, since a media player

is capable of beginning playback as the media file is being downloaded from an HTTP Web server. This fast-start playback relies on metadata located in the header of the media file (instead of the end of the media file in old formats) to retrieve the downloaded content in the local buffer.

Dynamic adaptive streaming over HTTP (DASH) (Stockhammer 2011) is a popular video delivery technique for conventional HTTP Web servers. DASH segments a media file into chunks of short intervals at multiple playback rates (encoding bit rates). Based on the media presentation description (MPD) file, which describes segment information (timing, URL, media characteristics such as video resolution and bit rates), a video client selects a bit rate to enable a fast session start-up and adaptively switches the bit rate of the next segment to request based on network conditions, device capabilities, user preferences, and so on. The bit rate adaptation is a key component to maintain high-quality video delivery with minimum stalls or rebuffering events during the playback.

To gauge user-perceived service quality, mean opinion score (MOS) is a widely used subjective QoE metric, which is usually expressed in a 5-point scale, where 5 equals *Excellent*, 4 equals *Good*, 3 equals *Fair*, 2 equals *Poor*, and 1 equals *Bad*. Subjective QoE depends on both picture distortions of video coding and transmission impairments. In Seshadrinathan et al. (2010), Seshadrinathan et al. used a four-parameter monotonic logistic function to fit objective video quality assessment (VQA) metrics to subjective difference mean opinion score (DMOS). A variety of VQA metrics that focus on the low-level video characteristics are evaluated, e.g., peak signal-to-noise ratio (PSNR), structural similarity (SSIM), video quality metric (VQM), and motion-based video integrity evaluation (MOVIE) index. Simpler video parameters are considered in the acceptability QoE model proposed in Song and Tjondronegoro (2014) for mobile video, such as quantization parameter, logarithm bit rate, frame rate, and content type. Instead of using the 5 or 11 scales of MOS, the acceptability QoE model adopts binary measures

that access general acceptability and pleasant acceptability of users for mobile video.

There are also QoE models that further incorporate network-related parameters. For instance, in Li et al. (2014), Li et al. carried out experiments with 30 participants to assess user-perceived QoE for video and audio clips delivered over the IEEE 802.11 wireless network in a lab environment. The subjective QoE quantified in a 5-point MOS scale is related to three objective metrics, including bit error probability, packet loss rate, and packet delay. It is found that the subjective QoE is sensitive to audio/video content although consistent with objective QoS metrics. Statistical difference-of-means tests show that the video with slower motion and fewer colors is likely to offer better delay tolerance.

Different from subjective assessment that directly involves tests with human users, a QoE model in objective assessment often uses a mathematical model to link MOS with objective QoE factors at different layers (Rengaraju et al. 2012; Reichl et al. 2013). This becomes a more viable approach to estimate the user-perceived QoE. Key performance indicators (KPIs) at the network level (e.g., throughput, packet loss, delay, and jitter) can capture quality at the network level (i.e., QoS), whereas quality at the application level is related to key quality indicators (KQIs), (e.g., start-up delay, freezing interruptions, and picture distortions for video streaming) (Rengaraju et al. 2012).

In G.1070 recommendation (ITU-T 2007) of ITU-T, the QoE model for video-telephony applications accounts for network parameters such as one-way end-to-end delay and packet loss, as well as video codec parameters, e.g., spatial resolution, frame rate, and encoding bit rate. In Khan et al. (2010), the heuristic MOS prediction model takes into account a network-level parameter, packet error rate, in addition to application-level parameters such as video bit rate and frame rate.

The logarithm scaling law in service quality perception is discussed in Reichl et al. (2013) from the perspectives of microeconomics and psychophysics. The authors also clearly distin-

guish between the quality degradation parameters like packet loss rate and resource-based stimulus parameters like transmission bandwidth and encoding bit rate. Unfortunately, the logarithm laws are only justified in the proposed QoE models for Web browsing and file download in the universal mobile telecommunications system (UMTS) with high-speed packet access (HSPA). The additive log-logistic model for networked video proposed in Zhang et al. (2013) can capture the relationship of visual quality against impairment-related features caused by both compression and transmission. In particular, the freezing duration due to packet loss is referred as one candidate metric, while a higher prediction accuracy can be achieved with more additional features.

Next, we introduce an example QoE model to give the readers a more concrete understanding. In Dobrian et al. (2011), Dorian et al. propose a number of video quality metrics based on user perception, including join time, buffering ratio, rate of buffering events, average bit rate, and rendering quality. The QoE model defined in De Vriendt et al. (2013) for adaptive video services involves average segment quality level and standard deviation of segment quality level. The QoE model proposed in Mok et al. (2011) further incorporates the impact of video freezes by taking into account the duration and frequency of video freezes. A comprehensive QoE model in Claeys et al. (2014) combines the two models in De Vriendt et al. (2013) and Mok et al. (2011). It incorporates *the average quality level, the standard deviation of the quality level, and the duration and frequency of video freezes* and quantifies MOS by

$$MOS = \max(5.67 \cdot \mu - 6.72 \cdot \sigma - 4.95 \cdot \phi + 0.17, 0)$$

where

$$\mu = \frac{\sum_{k=1}^K \frac{Q_k}{N}}{K}, \quad \sigma = \sqrt{\frac{\sum_{k=1}^K \left(\frac{Q_k}{N} - \mu \right)^2}{K - 1}}$$

$$\phi = \frac{7 \cdot \max\left(\frac{\ln F_{freq}}{6} + 1, 0\right) + \frac{\min(F_{avg}, 15)}{15}}{8}.$$

Here, N denotes the total number of quality levels and K denotes the total number of segments in the video. Moreover, Q_k is the quality level requested for the k th segment, F_{freq} is the ratio of the number of video freezes to the number of segments, and F_{avg} is the average duration of all freezes. As seen, the QoE captures quality metrics in various aspects and provides a holistic assessment of user-perceived quality.

Key Applications

QoE assessment models are fundamental in network resource management for multimedia service provisioning in order to enhance user experience.

We still take video services as an example to introduce a key application of QoE models. It is known that wireless networks are highly varying due to channel fading and user mobility. To counter against the high dynamics and deliver high-quality video, adaptive bit rate streaming (ABS) is an effective technique for HTTP-based video delivery. ABS intends to provide users with the best available video quality in corresponding network condition by adopting bit rate adaptation schemes. Traditionally, many bit rate adaptation schemes (Akhshabi et al. 2011; Thang et al. 2012; Miller et al. 2012; Liu et al. 2011) aim at the improvement of video QoS such as throughput, packet loss, starvation probability, freezing duration and rate, etc. Here, freezing and starvation mean that the playback buffer at the end user runs below the smallest video decoding block such that the playback is interrupted.

With an appropriate QoE model, ABS can target at QoE improvement, i.e., the improvement of overall user-perceived quality. A QoE model can capture the effects of a variety of objective metrics on user perception. When the QoE model is used to guide the bit rate adaptation, the holistic perception of QoE rather than some isolated QoS metrics can be improved. For example, in Bentalab et al. (2016), the authors consider a QoE model that utilizes the most influential metrics on end-user satisfaction including start-up delay, the number of stalls, average video quality, and the number of video quality switches. Accordingly,

the optimal bit rate for the next segment to be downloaded is determined dynamically to maximize the QoE of each video client.

Cross-References

- [QoS in Indoor Localization](#)
- [QoS-Aware MAC](#)
- [Quality of Service in IEEE 802.11 Networks](#)
- [Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes](#)

References

- Akhshabi S, Begen AC, Dovrolis C (2011) An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP. In: Proceedings of ACM conference on multimedia systems, Santa Clara, pp 157–168
- Bentalab A, Ozyegin ACB, Zimmermann R (2016) SDNDASH: improving QoE of HTTP adaptive streaming using software defined networking. In: Proceedings of ACM multimedia conference, Amsterdam, pp 1296–1305
- Brunnström K, et al (2013) Qualinet white paper on definitions of quality of experience. European network on quality of experience in multimedia systems and services (COST Action IC 1003), Lausanne
- Cisco (2017) Cisco visual networking index: global mobile data traffic forecast update, 2016–2021
- Claeys M, Latré S, Famaey J, De Turck F (2014) Design and evaluation of a self-learning HTTP adaptive video streaming client. *IEEE Commun Lett* 18(4):716–719
- De Vriendt J, De Vleeschauwer D, Robinson D (2013) Model for estimating QoE of video delivered using HTTP adaptive streaming. In: Proceedings of IFIP/IEEE international symposium on integrated network management (IM), Ghent, pp 1288–1293
- Dobrian F, Sekar V, Awan A, Stoica I, Joseph D, Ganjam A, Zhan J, Zhang H (2011) Understanding the impact of video quality on user engagement. *ACM SIGCOMM Comput Commun Rev* 41(4):362–373
- ITU-T (2007) ITU-T recommendation G.1070, opinion model for video-telephony applications
- ITU-T (2008) ITU-T recommendation G.100/P.10, vocabulary for performance and quality of service, amendment 2: new definitions for inclusion in recommendation P.10/G.100
- Khan A, Sun L, Jammeh E, Ifeakor E (2010) Quality of experience-driven adaptation scheme for video applications over wireless networks. *IET Commun* 4(11):1337–1347
- Li X, Dong K, Song W, Nickerson B (2014) QoE evaluation of multimedia transmission over wireless networks. In: Proceedings of international wireless communications and mobile computing conference (IWCMC), Nicosia

- Liu C, Bouazizi I, Gabbouj M (2011) Rate adaptation for adaptive HTTP streaming. In: Proceedings of ACM conference on multimedia systems, Santa Clara, pp 169–174
- Miller K, Quacchio E, Gennari G, Wolisz A (2012) Adaptation algorithm for adaptive streaming over HTTP. In: Proceedings of international packet video workshop, Munich, pp 173–178
- Mok RK, Chan EW, Chang RK (2011) Measuring the quality of experience of HTTP video streaming. In: Proceedings of IFIP/IEEE international symposium on integrated network management (IM), Dublin, pp 485–492
- Reichl P, Tuffin B, Schatz R (2013) Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience. *Telecommun Syst* 52(2):587–600
- Rengaraju P, Lung CH, Yu FR, Srinivasan A (2012) On QoE monitoring and E2E service assurance in 4G wireless networks. *IEEE Wirel Commun* 19(4):89–96
- Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. *IEEE Trans Image Process* 19(6):1427–1441
- Song W, Tjondronegoro DW (2014) Acceptability-based QoE models for mobile video. *IEEE Trans Multimed* 16(3):738–750
- Stockhammer T (2011) Dynamic adaptive streaming over HTTP: standards and design principles. In: Proceedings of ACM conference on multimedia systems, Santa Clara, pp 133–144
- Thang TC, Ho QD, Kang JW, Pham AT (2012) Adaptive streaming of audiovisual content using MPEG DASH. *IEEE Trans Consum Electron* 58(1):78–85
- Wang Y, Li P, Jiao L, Su Z, Cheng N, Shen X, Zhang P (2017) A data-driven architecture for personalized QoE management in 5G wireless networks. *IEEE Wirel Commun* 24(1):102–110
- Zhang F, Lin W, Chen Z, Ngan KN (2013) Additive log-logistic model for networked video quality assessment. *IEEE Trans Image Process* 22(4):1536–1547

QoE Measurement and Assessment of Video Streaming

Xinggong Zhang and Zongming Guo
WangXuan Institute of Computer Technology,
Peking University, Beijing, China

Synonyms

[HTTP adaptive streaming](#); [Internet video streaming](#); [QoE assessment and measurement](#); [Quality of experience \(QoE\)](#)

Definition

MOS	Mean opinion score
PSNR	Peak signal to noise ratio
QoE	Quality of experience
QoS	Quality of services
SSIM	Structural similarity

History Background

Past few years have witnessed the booms of Internet video. The online video service providers, such as YouTube, Amazon, Hulu from USA and Youku, Tencent, Toutiao from China are becoming the main players in the market of video entertainment (CNNIC 2017; Hossfeld et al. 2011). Mobile phones, over-the-top (OTT) devices, and online streaming are substituting televisions as the new favorable ways for the generation born after 1980. It is necessary for the providers to assess the service quality.

Quality assessment is firstly proposed for digital television, to evaluate the quality of video coding and transmission. Many metrics have been proposed (Seufert et al. 2015), such as the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM), and the subjective metric of the mean opinion score (MOS). The video quality is measured by comparing the original videos with the compressed video, or human subjects are asked to rate the quality.

Quality of service (QoS) is defined by ITU (ITU 2012) to measure the performance of network, not the actual experience of user. The common QoS metrics are throughput, packet losses, delay, jitter, etc.

Online video streaming is an end-to-end video delivery and playback. Its quality depends on video coding and on network conditions as well. Mostly, its quality is assessed by quality of experience (QoE), a user-centric metric that measures the performance subjectively perceived by the user.

The impact factors on QoE are classified into three categories:

1. Video level: video quality (PSNR), frame rate, and resolution

2. Application level: start-up delay, bit rate, stall/rebuffering, and rate oscillation
3. Context level: screen size, browser/player, and screen distance

Key Applications

To capture user's QoE, the streaming methods should be known at first since various applications have different requirements on QoE (Juluri et al. 2013). The metrics to measure the QoE have two types: objective metrics and subjective metrics. The metrics can be measured on either network side or client side by different measurement tools.

Streaming method is one of the most important factors affecting QoE. The widely used streaming technologies include real-time streaming (RTS) (Mazhar et al. 2018), HTTP progressive downloading (HPD), and HTTP adaptive streaming (HAS) (Liu et al. 2013; Zhang et al. 2012). All of them are able to enable users to start the playback once the part of the video is downloaded. However, due to their different transmission technologies, their quality impairments are not same. For example, RTS is mainly used in low-latency interactive applications, such as live streaming and video chatting. It is not only sensitive to video quality, but also to round-trip delay.

QoE metrics are used to assess the quality of video streaming. It can be classified into two categories: (Eswara et al. 2018) metrics and (Ghadiyaram et al. 2019). Objective metrics are the metrics which can be quantified with a measurement tool, such as bit (Xu et al. 2019) and delay. These metrics are objective and easy to be measured. However, they have only indirect impacts on users' experience with the service. Subjective metrics are the direct QoE feedbacks from users. Users rate the video service on a controlled environment. However, subjective metrics are susceptible to bias because users' QoE could be varied from one subject to another.

The techniques to measure QoE are also important for video streaming. Video streaming is an end-to-end service. There are multiple

parties participating in it, such as content providers, content-distribution-network (CDN) providers (Mangla 2019), and (Krishnamoorthi 2017). They view the end-to-end streaming from different perspectives, thus the QoE can be measured with the different measurement methodologies and tools.

References

- CNNIC (2017) The 40th CNNIC statistical survey report on internet development in China
- Eswara N et al (2018) A continuous QoE evaluation framework for video streaming over HTTP. *IEEE Transactions on circuits and systems for video technology* 28(11):3236–3250
- Ghadiyaram D et al (2019) A subjective and objective study of stalling events in mobile streaming videos. *IEEE Transactions on Circuits and Systems for Video Technology* 29(1):183–197
- Hossfeld T et al (2011) Quantification of YouTube QoE via crowdsourcing, 2011 IEEE international symposium on multimedia, Dana Point, 2011, pp. 494–499
- ITU-T (2012) Vocabulary for performance and quality of service: recommendation P.10/G.100-Amendment 3
- Juluri P et al (2013) Viewing YouTube from a metropolitan area: what do users accessing from residential ISPs experience, 2013 IFIP/IEEE international symposium on integrated network management (IM 2013), Ghent, pp. 589–595
- Krishnamoorthi V (2017) BUFFEST: Predicting buffer conditions and real-time requirements of HTTP(S) adaptive streaming clients. *Proc ACM MMSys*, pp. 76–87
- Liu et al (2013) A study on quality of experience for adaptive streaming service, 2013 IEEE International Conference on Communications Workshops (ICC), Budapest, 2013, pp. 682–686
- Mangla T (2019) Using session modeling to estimate HTTP-based video QoE metrics from encrypted network traffic. *IEEE Transactions on Network and Service Management*. 16(3):1086–1099
- Mazhar MH et al (2018) Real-time video quality of experience monitoring for HTTPS and QUIC, *IEEE INFOCOM 2018 – IEEE Conference on Computer Communications*, Honolulu, 2018, pp. 1331–1339
- Seufert M et al (2015) A survey on quality of experience of HTTP adaptive streaming. *IEEE Commun Surv Tutorials* 17(1):469–492
- Xu Z et al (2019) QoE-driven adaptive K-push for HTTP/2 live streaming. *IEEE Transactions on Circuits and Systems for Video Technology* 29(6):1781–1794
- Zhang X et al (2012) A control-theoretic approach to rate adaptation for dynamic HTTP streaming, 2012 visual communications and image processing, San Diego, 2012, pp. 1–6

QoE Metrics

- [QoE Assessment Models](#)

QoE Models

- [QoE Assessment Models](#)

QoS Enhancements in IEEE 802.11e

- [Quality of Service in IEEE 802.11 Networks](#)

QoS in Indoor Localization

Xiaohua Tian
Shanghai Jiao Tong University, Shanghai, China

Synonyms

[QoS in indoor positioning](#); [QoS in wireless indoor localization](#); [QoS in wireless indoor positioning](#)

Definitions

Wireless indoor localization systems locate the target's position in indoor environment, where the GPS signal is obstructed by walls of buildings. Current wireless indoor localization systems can be implemented with different techniques like Wi-Fi, Bluetooth Low Energy (BLE), radio frequency identification (RFID), ultra-wideband (UWB), and so on. The quality of service (QoS) in each indoor localization system is measured with accuracy, time delay, computational complexity, and availability.

Historical Background

How to obtain the position of the target precisely is always an important problem to solve, which plays a special role no matter in military or people's daily life. The most famous and applicable positioning system is Global Positioning System (GPS) built with huge amount of resources by the United States in the 1970s. The GPS system provides critical positioning capabilities to military, civil, and commercial users around the world with relatively high QoS. However, because the radio signal is obstructed by solid object like walls, GPS is not functional in indoor environment. Thus, the research of indoor localization is an urgent need which possesses a huge potential.

In the early stage of the 1980s, technologies like infrared or ultrasonic wave is adopted, such as the Active Badge and Active Bat system which are developed by Cambridge University. These systems can achieve high accuracy in indoor environment but have many limitations like coverage area and cost. In recent years, wireless radio frequency (RF) indoor localization, which can utilize radio frequency such as Wi-Fi and Bluetooth, attracts people's attention. The low-cost and easy-deployment properties of RF indoor localization system are beneficial to the large-scale deployment and also enhance the QoS in indoor localization.

In 2000, Microsoft published RADAR System (Bahl and Padmanabhan 2000), which is the world's first Wi-Fi signal strength-based indoor localization system but limited in accuracy and availability. Based on RADAR System, many researches have been done, and the QoS in indoor localization is enhanced gradually. Specifically, Youssef proposes the Horus system (Youssef and Agrawala 2005), which identifies different causes for the wireless channel variations and addresses them to achieve high accuracy. The EZ localization algorithm (Chintalapudi et al. 2010) models the constraints of environments by analyzing wireless propagation, to save pre-deployment efforts. Zee system (Rai et al. 2012) makes the calibration zero-effort, by enabling training data to be crowdsourced without any explicit effort on the part of users. Wen presents a general prob-

abilistic model (Wen et al. 2015) to explore the fundamental limits of Received Signal Strength (RSS) fingerprinting-based indoor localization. Mei (Wang et al. 2016b) evaluates how the temporal correlation of the RSS can influence the reliability of location estimation. Last but not the least, (Wang and Katabi 2013) introduces a fine-grained RFID positioning system that is robust to multipath and non-line-of-sight scenarios.

The researches above are all based on the Received Signal Strength Indicator (RSSI). However, RSSI suffers from dramatic performance degradation in complex situations due to multipath fading and temporal dynamics. Thus, nowadays channel state information (CSI) has attracted many researchers' attentions, and some pioneer works have demonstrated submeter- or even centimeter-level accuracy. For example, Phaser system (Gjengset et al. 2014), with the adoption of CSI, enables phased array signal processing for commodity Wi-Fi APs. SpotFi (Kotaru et al. 2015) accurately computes the angle of arrival (AOA) of multipath components by the measurable changes in CSI. LiFS (Wang et al. 2016a) uses CSI to realize device-free localization, and Chronos (Vasisht et al. 2016) enables a single Wi-Fi access point to locate clients with CSI.

Foundations

A typical wireless RF indoor localization system is consisted of three parts: measurement technique, selected feature, and mapping method. A basic localization procedure of wireless indoor localization is like the following: First, the device with the specific measurement technique will measure some monitoring information from the indoor environment. Then, the device extracts some features from the monitoring information. Finally, with the mapping method, we can map these features into real positions.

Specifically, the first part is measurement technique. There are four common types of measurement technique: Wi-Fi, RFID, BLE, and UWB. Here are the details of the four techniques:

- *Wi-Fi*: Wi-Fi is a technology that allows devices to be connected into a wireless local area network (WLAN). It is very popular and widely used in homes, hotels, airports, and shopping malls, which makes Wi-Fi one of the most eye-catching wireless technique in the field of indoor localization. With a high availability, the QoS of Wi-Fi-based system is also largely improved. Typically, a Wi-Fi system consists of some fixed access points (APs), which are largely deployed in easy-to-access location, and the network administrator knows the distribution of the APs. Therefore, Wi-Fi-based localization systems can easily reuse the existing WLAN infrastructures in public or private indoor environments, which lower the cost of deployment for localization.
- *BLE*: Bluetooth is a wireless technology standard that enables short-distance data exchange between devices. Many types of devices like mobile phones, laptops, and tablets are implanted with Bluetooth, and many connected devices can form a Bluetooth cluster. In Bluetooth-based localization systems, the position of devices can be located by the effort of other mobile devices in the same Bluetooth cluster. BLE is a new technique based on Classic Bluetooth, aiming at providing considerably low power consumption and cost while maintaining a similar communication range. With the availability increasing, the QoS of BLE is also enhanced.
- *RFID*: RFID is a communication technology that can identify specific targets through radio signal and then reads, writes, or stores data into a RF-compatible integrated circuit. RFID-based localization systems are commonly deployed in complex indoor environments, like offices and hospitals. There are two kinds of RFID: passive RFID and active RFID. The tags with passive RFID are usually small and inexpensive, but short in accuracy, which leads to lower QoS. On the contrary, with higher-cost and energy-consuming properties, active RFID could achieve better accuracy and shorter time delay, leading to high QoS.

- **UWB:** UWB is a carrier-free communication technology that can transmit data with nanosecond or picosecond non-sinusoidal pulses. By sending extremely low power signals over a wide spectrum, the UWB can achieve high data transmission rate from 100 Mbit/s to 100 Gbit/s in a range of 10 m. For UWB pulses having a duration less than 1 ns, it is possible for UWB signal to pass through the walls and reduce the multipath distortion. With submeter or even centimeter accuracy, UWB-based localization system can achieve high QoS. However, the cost of UWB device is expensive, which is short in applicability.

The second part is a selected feature. There are generally three kinds of it: power features, time features, and angle features. Such features can be extracted from the monitoring information, which is collected by the measurement technique. Specifically, RSSI and CSI are the most common monitoring information categories.

RSSI is a kind of signal strength information collected from the MAC layer of the mobile device. It can characterize the attenuation of radio signals after propagation. RSSI is a typical power feature and accessible with most wireless technique. For RSSI is easy to monitor and handily access, it has been widely adopted in both geometric mapping and fingerprinting localization systems. RSSI-based indoor localization can achieve meter-level high accuracy in some simple environments. However, the QoS of RSSI-based indoor localization in complex indoor environment is unsatisfactory, due to multipath fading and temporal dynamics. Generally speaking, traditional power features like RSSI could only remain room-level accuracy in complex environments.

CSI is a kind of channel response information collected from the PHY layer of the mobile device, which can show the channel response in 802.11 a/g/n. For CSI is collected from lower layer than RSSI, there are more fine-grained information and features contained in CSI. Various power, time, and angle features could

be extracted from CSI. Particularly, CSI is able to characterize multipath problem. Compared to RSSI, CSI holds more potential for better QoS in real complex indoor environments.

Specifically, from the channel response, the amplitude of signals in time domain or frequency domain could be extracted as the power features. Typical time features extracted from CSI include time of arrival (TOA) and time difference of arrival (TDOA). For example, Chronos (Vasisht et al. 2016) utilizes different channels of single Wi-Fi to extract TOA, which achieves high QoS of decimeter accuracy. Angle features like angle of arrival (AOA) provide a different thought for geometric mapping, which is utilized in SpotFi (Kotaru et al. 2015) with the help of subcarriers' information.

The usability of channel response is much lower than RSSI, for CSI is rather new and previously it could only be measured by professional equipment. However, with the development of Wi-Fi and orthogonal frequency-division multiplexing (OFDM), the situation has been greatly improved. Nowadays in 802.11 a/g/n environment, channel response could be easily extracted from off-the-shelf OFDM receivers. In recent years, numerous efforts have been paid on CSI to improve the QoS of wireless indoor localization.

The last part is mapping method, which is adopted to convert physical measurements into real positions. Specifically, geometric mapping and fingerprinting are the most common mapping methods.

In geometric mapping, distance and direction to the reference points could be the intermediate geometric parameters. These parameters could be derived from the physical measurement. Then, with certain geometric algorithm like triangulation, such parameters could be converted into real positions. Specifically, according to the geometric properties of triangles, there are four common methods of triangulation: RSS, TOA, TDOA, and AOA. RSS is easy to get, however, with the property of fluctuation in indoor environment, leading to the decrease in QoS. TOA can accurately derive the distance. But common devices are not capable to get the precise TOA. TDOA

utilizes the difference of arrival time between receivers, which is easier to implement compared with TOA. AOA determines target's position with the help of angles between transceivers.

Different with geometric mapping, fingerprinting is a pattern matching method. There are two phases in fingerprinting: offline training phase and online position determination phase. In the offline training phase, the basic idea is to pre-collect the selected features of all the possible locations in the area and build a fingerprint database to store the collected data related to the locations. Then in the online position determination phase, the location-related data of the position to be located is measured and sent to the fingerprint database. The database will return the location corresponding to the best-fitted fingerprint data. Generally speaking, fingerprinting method costs more effort, for it needs to build and frequently update the fingerprint database.

Compared to geometric mapping, fingerprinting method lowers the requirement of physical measurement. However, to avoid confusion and ensure the QoS of fingerprinting database, the fingerprint data of the same location collected at different time stamps should be close to each other.

Here is a typical example of wireless indoor localization. SpotFi (Kotaru et al. 2015) is a CSI-based Wi-Fi indoor localization system. There are three steps of SpotFi: First, use CSI to estimate the AOA and time of flight (TOF) of different multipath components of a target's signal arriving at the AP. Then, estimate the likelihood that each AOA and TOF pair is the one corresponding to the direct path between the AP and the target without any reflections. Finally, use above information to calculate the most likely location of the target that could have produced the observed RSSI and estimated AOA. The experiments in a multipath-rich indoor environment show that SpotFi has a remarkable QoS, which achieves a median accuracy of 40 cm and is robust to indoor hindrances such as obstacles and multipath.

Different from the typical wireless indoor localization system, there is a new trend of wireless indoor localization called passive

detection (device-free localization). Passive detection means that the system can detect the presence of or even locate a target by the deployed wireless monitors, while there is no any detectable device on the target. The principle of passive detection is that when a target object presents and moves in the area of interest, there must be some certain changes of the received signal features on the wireless monitor side. By analyzing the changes and adopting the proper algorithm, the device-free target could be located.

Key Applications

Wireless indoor positioning systems have been designed to provide location information of persons or devices in indoor environment. It enables devices to determine the target's position and makes the position of the target available for position-based services such as navigating, tracking, or monitoring in indoor environment.

Cross-References

- [QoE Assessment Models](#)
- [Quality of Service in IEEE 802.11 Networks](#)

References

- Bahl P, Padmanabhan VN (2000) Radar: an in-building RF-based user location and tracking system. In: Proceedings of the nineteenth annual joint conference of the IEEE computer and communications societies, INFOCOM 2000, vol 2. IEEE, pp 775–784
- Chintalapudi K, Padmanabha Iyer A, Padmanabhan VN (2010) Indoor localization without the pain. In: Proceedings of the sixteenth annual international conference on mobile computing and networking. ACM, pp 173–184
- Gjengset J, Xiong J, McPhillips G, Jamieson K (2014) Phaser: enabling phased array signal processing on commodity wifi access points. In: Proceedings of the 20th annual international conference on mobile computing and networking. ACM, pp 153–164
- Kotaru M, Joshi K, Bharadia D, Katti S (2015) SpotFi: decimeter level localization using wifi. In: ACM SIGCOMM computer communication review, vol 45. ACM, pp 269–282

- Rai A, Chintalapudi KK, Padmanabhan VN, Sen R (2012) Zee: zero-effort crowdsourcing for indoor localization. In: Proceedings of the 18th annual international conference on mobile computing and networking. ACM, pp 293–304
- Vasisht D, Kumar S, Katabi D (2016) Decimeter-level localization with a single wifi access point. In: NSDI, pp 165–178
- Wang J, Katabi D (2013) Dude, where's my card? RFID positioning that works with multipath and non-line of sight. ACM SIGCOMM Comput Commun Rev 43(4):51–62
- Wang J, Jiang H, Xiong J, Jamieson K, Chen X, Fang D, Xie B (2016a) Lifis: low human-effort, device-free localization with fine-grained subcarrier information. In: Proceedings of the 22nd annual international conference on mobile computing and networking. ACM, pp 243–256
- Wang M, Zhang Z, Tian X, Wang X (2016b) Temporal correlation of the RSS improves accuracy of fingerprinting localization. In: IEEE INFOCOM 2016-the 35th annual IEEE international conference on Computer communications. IEEE, pp 1–9
- Wen Y, Tian X, Wang X, Lu S (2015) Fundamental limits of RSS fingerprinting based indoor localization. In: 2015 IEEE conference on computer communications (INFOCOM). IEEE, pp 2479–2487
- Youssef M, Agrawala A (2005) The horus wlan location determination system. In: Proceedings of the 3rd international conference on mobile systems, applications, and services. ACM, pp 205–218

QoS Provisioning at the MAC Layer

► QoS-Aware MAC

QoS Provisioning in IEEE 802.11

► Quality of Service in IEEE 802.11 Networks

QoS-Aware Hybrid Medium Access Control (MAC)

► Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs

QoS-Aware MAC

Hangguan Shan
Zhejiang University, Hangzhou, China

QoS in Indoor Positioning

► QoS in Indoor Localization

QoS in Wireless Indoor Localization

► QoS in Indoor Localization

QoS in Wireless Indoor Positioning

► QoS in Indoor Localization

Synonyms

Providing QoS support at the distributed wireless MAC protocol; QoS-capable MAC; QoS provisioning at the MAC layer

Definitions

In distributed wireless networks, a QoS-aware medium access control (MAC) protocol is used not only to address the contention and collision problem among mobile or fixed users sharing the same physical medium and the improvement of channel utilization but also to provide some QoS guarantees for users in, for example, terms of minimum throughput, maximum frame delay,

maximum variation of frame delay (jitter), and maximum frame loss ratio.

Historical Background

In general, the QoS-aware distributed MAC protocols evolve from the MAC protocols only supporting best-effort services. The first MAC, ALOHA (or pure ALOHA) (Abramson 1970), started out in Hawaii in the early 1970s, is a random access protocol proposed for packet radio networks. ALOHA and its improved version slotted ALOHA (Roberts 1975), as well as many early MAC protocols built based on Carrier Sense Multiple Access (CSMA) mechanism such as Multiple Access with Collision Avoidance (MACA) (Karn 1990) and Floor Acquisition Multiple Access (FAMA) (Fullmer and Garcia-Luna-Aceves 1997), aim at increasing network throughput by reducing collisions and/or solving hidden terminal and exposed terminal problems (Cheng et al. 2006).

However, as heterogeneous traffic emerge, QoS support with service differentiation at the MAC layer becomes indispensable. For example, delay-sensitive traffic such as voice may need a higher priority than the traffic such as data. To this end, both types of MAC protocols, enabling statistical priority and strict priority, were proposed. The Enhanced Distributed Channel Access (EDCA) function of IEEE 802.11e (IEEE Std 2007) and the Black-Burst (BB) contention scheme (Sobrinho and Krishnakumar 1999) are two well-known examples, belonging to each of the aforesaid types.

To achieve fine-grain QoS guarantee beyond priority-based MAC protocols, newly developed MAC protocols (e.g., distributed PRMA (Jiang et al. 2002)) also integrate with the function of resource reservation. With frame or slot reservation, it is possible to guarantee strict delay bound for voice packet, which is difficult if contention is unavoidable for every voice packet.

It is noteworthy that different wireless networks can have quite diverse QoS requirements. Furthermore, the QoS requirements and the tailored MAC protocols of a wireless network

also evolve with its applications and hardware advancements. A typical example is wireless sensor network (WSN). In the early stage, efficient data delivery (so throughput) was not its first priority, and more considerations had been given to how to extend the lifetime of the network effectively (Huang et al. 2013). However, the availability of low-cost hardware and rapid development of tiny cameras and microphones have enabled a new class of WSNs: multimedia or visual wireless sensor networks, which bring forward new QoS requirements of MAC on delay-bounded and reliable data delivery (Yigitel et al. 2011).

Foundations

In general, QoS provisioning at the MAC layer could be designed by following two principal approaches: prioritized, based on differentiated services (DiffServ) architecture, and parameterized, based on integrated service (IntServ) architecture (Natkaniec et al. 2013). DiffServ is a coarse-grained, class-based mechanism for traffic management. It operates on the principle of traffic classification, where each frame is managed and served at nodes along a delivery path according to its traffic class, rather than differentiating network traffic based on the requirements of an individual flow. In contrast, IntServ is a fine-grained, flow-based mechanism, in which to satisfy a flow's strict QoS requirements, deterministic capacity has to be reserved along the delivery path in advance. However, in a distributed wireless network, as the transmission is usually quality-unreliable and prone to collisions (thus somewhat capacity-unpredictable), DiffServ-based QoS provisioning is found more reasonable for a QoS-aware distributed MAC protocol.

In the existing distributed MAC protocols, the following mechanisms are employed frequently to provide QoS in distributed wireless networks, namely, backoff differentiation, InterFrame space differentiation, jamming, frame aggregation, frame manipulation, reservation, and alternating

contention period (CP)/contention-free period (CFP) (Natkaniec et al. 2013).

Backoff differentiation supports traffic differentiation by extending the basic function of backoff, assigning different contention window parameters for different traffic classes. For example, as compared with the basic Distributed Coordination Function in IEEE 802.11a or b, the EDCA function of IEEE 802.11e defines four access categories (ACs): Voice (Vo), Video (Vi), Best Effort (BE), and Background (BK) (IEEE Std 2007). Each AC has its own contention window minimum and maximum values ($CW_{\min}[AC]$ and $CW_{\max}[AC]$). To give preference to high-priority traffic, the contention window of high-priority traffic should be smaller than that of low-priority traffic.

InterFrame space differentiation is usually used together with backoff differentiation. Taking the EDCA function of IEEE 802.11e as an example, instead of using Distributed InterFrame Space (DIFS) for all traffic class, the backoff counter of traffic class AC may begin decrementing after the Arbitrary InterFrame Space (AIFS) associated with its own AC, AIFS[AC] (IEEE Std 2007). So, differentiated access control is provided by letting high-priority traffic have shorter AIFSs. However, neither backoff differentiation nor InterFrame space differentiation can provide strict priority differentiation. It is possible that a node with high-priority traffic fails in channel access when contending with a node with low-priority traffic.

To maintain better traffic differentiation, jamming, sometimes named black bursts (Sobrinho and Krishnakumar 1999) or busy tones (Haas and Deng 2002), can be utilized. The basic idea of jamming is to let a contender send a jamming signal of which length is proportional to the priority; thus other contenders will quit channel access contention if detecting a longer jamming signal. To ensure a good priority differentiation resolution, the setting of the jamming signal length should take into account the network area or the largest propagation delay in the network. However, jamming may also increase energy consumption and/or consume extra bandwidth, due to the transmission of the jamming signal itself.

To improve channel utilization, reducing the number of channel contentions by frame aggregation is exploited by multiple MAC protocols. Priority Unavoidable Multiple Access (PUMA) (Natkaniec and Pach 2002) is the first QoS-aware MAC protocol that uses the approach but calls it packet-train mechanism. In IEEE 802.11, this mechanism is referred to as Transmission Opportunity (TXOP) (IEEE Std 2007). The standard defines TXOP as a time interval during which a node has right to transmit frame(s) in sequence.

To increase network throughput and provide better QoS, frame manipulation is adopted in several MAC protocols. In specific, when a frame waiting in a MAC-layer queue is expired for delay violation or useless due to a loss of other dependent frame(s) at the receiver, the frame can be dropped at the transmitter side directly (Pal et al. 2002). The goal of changing frame priority is to increase the channel access probability of a specific frame, thus finishing delivery on time or before deadline (Natkaniec and Pach 2002).

If sensitive real-time traffic has periodic frames, performing channel reservation could be a good choice, to not only reduce its channel contention overhead but also guarantee its delay performance. In the literatures, channel reservation could be employed based on either Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) or Time-Division Multiple Access (TDMA). The Multiple Access Collision Avoidance with Piggyback Reservations (MACA/PR) protocol (Lin and Gerla 1997) and the Soft Reservation Multiple Access with Priority Assignment (SRMA/PA) protocol (Ahn et al. 2000) are two of this type of typical MAC protocols, built based on CSMA/CA and TDMA, respectively.

When heterogeneous traffic, e.g., both delay-sensitive and best-effort traffic, coexist in a distributed wireless network, alternating CP and CFP provides opportunities to serve them differentially and with better QoS provisioning. In general, the Point Coordination Function (PCF) defined in IEEE 802.11 is a representative MAC protocol following this approach (IEEE Std 2007). However, PCF was designed for infrastructure network configuration, requiring

a coordinator (called the Access Point, AP) to poll other nodes for data transmission in CFP, and it suffers from preventing stations in another nearby network from transmitting competing traffic. In case of no AP available, dynamically creating the infrastructure and choosing node as the coordinator are considered in MAC protocols such as the Mobile Point Coordinator for MAC (MPC-MAC) (You et al. 2002).

Key Applications

QoS-aware MAC protocol is a fundamental component of every distributed wireless network where delay-sensitive traffic or heterogeneous traffic should be served.

Cross-References

- [Multiple Access Technique for Cellular Wireless Networks](#)
- [Quality of Service in IEEE 802.11 Networks](#)
- [Random Access Technology in Cellular Networks](#)

References

- Abramson N (1970) The ALOHA system: another alternative for computer communications. In: AFIPS conference proceedings of fall joint computer conference
- Ahn CW, Kang CG, Cho YZ (2000) Soft reservation multiple access with priority assignment (SRMA/PA): a novel MAC protocol for QoS-guaranteed integrated services in mobile ad-hoc networks. In: Proceedings of the IEEE VTC-Fall'00
- Cheng HT, Jiang H, Zhuang W (2006) Distributed medium access control for wireless mesh networks. *Wirel Commun Mobile Comput* 6:845–864
- Fullmer CL, Garcia-Luna-Aceves JJ (1997) Solutions to hidden terminal problems in wireless networks. In: Proceedings of the of ACM SIGCOMM
- Haas ZJ, Deng J (2002) Dual busy tone multiple access (DBTMA)-a multiple access control scheme for ad hoc networks. *IEEE Trans Commun* 50:975–985
- Huang P, Xiao L, Soltani S, Mutka MW, Xi N (2013) The evolution of MAC protocols in wireless sensor networks: a survey. *IEEE Commun Surv Tutor* 15: 101–120
- IEEE Std (2007) IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks specific requirements part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications
- Jiang S, Rao J, He D, Ling X, Ko C (2002) A simple distributed PRMA for MANETs. *IEEE Trans Veh Technol* 51:293–305
- Karn P (1990) MACA-A new channel access method for packet radio. In: Proceedings of the the ARRL/CRRL amateur radio 9th computer networking conference
- Lin C, Gerla M (1997) Asynchronous multimedia multihop wireless networks. In: Proceedings of the IEEE INFOCOM'97, sixteenth annual joint conference of the IEEE computer and communications societies
- Natkaniec M, Pach AR (2002) PUMA – a new channel access protocol for wireless LANs. In: 5th international symposium on wireless personal multimedia communications
- Natkaniec M, Kosek-Szott K, Szott S, Bianchi G (2013) A survey of medium access mechanisms for providing QoS in ad-hoc networks. *IEEE Commun Surv Tutor* 15:592–619
- Pal A, Dogan A, Ozguner F (2002) MAC layer protocols for real-time traffic in ad-hoc wireless network. In: Proceedings of the international conference on parallel processing
- Roberts LG (1975) ALOHA packet system with and without slots and capture. *Comput Commun Rev* 5:28–42
- Sobrinho J, Krishnakumar A (1999) Quality-of-service in ad hoc carrier sense multiple access wireless networks. *IEEE J Sel Areas Commun* 17:1353–1368
- Yigitel MA, Incel OD, Ersoy C (2011) QoS-aware MAC protocols for wireless sensor networks: a survey. *Comput Netw* 55:1982–2004
- You T, Hassanein H, Moutfah HT (2002) Infrastructure-based MAC in wireless mobile ad-hoc networks. In: Proceedings of the 27th annual IEEE conference on local computer networks

QoS-Capable MAC

- [QoS-Aware MAC](#)

Quality of Experience

- [Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks](#)

Quality of Experience (QoE)

► QoE Measurement and Assessment of Video Streaming

Quality of Experience for Wireless Video Streaming

Jie Li, Ransheng Feng, and Qiyue Li
School of Computer and Information, Hefei
University of Technology, Hefei, Anhui, China

Synonyms

Quality of Experience(QoE) for wireless video transmission; Quality of Service (QoS)for wireless video streaming

Definitions

Quality of Experience for wireless video streaming, which measures the quality that subjectively perceived by the end users, is able to directly and accurately reflect the subjective feelings of viewers when users watch streaming video over the wireless network.

Historical Background

Video streaming services are becoming more and more popular these days. For multimedia streaming service providers, a good video quality may lead to an increase in customer subscriptions. On the contrary, a poor video quality may result in subscriber churn. Thus, it becomes more important for the video streaming service providers to increase the user perceptual video quality. However, it is not easy to measure the user perceptual quality of a video. Classical QoS (Jin and Nahrstedt 2004) metrics in mobile networks are blocking rates for real-time traffic and average user throughput for elastic one and operators

dimension their networks for satisfying targets on those metrics.

Regarding objectively measuring the performance of a video transmission system from video server to users, QoS is considered as the most successful system quality strategy. It is widely applied in video wireless transmission system as an optimization goal. However, QoS may fail to capture the quality degradation of the viewing experience caused by factors like compression artifacts. Compared to QoS, QoE (Zhao et al. 2017), which measures the quality that subjectively perceived by the end users, is able to directly and accurately reflect the subjective feelings of the viewer.

In early works, researchers were trying to increase user perceptual video quality by appropriately selecting QoS parameters (such as video compression optimization (Sullivan and Wiegand 1998; Chen and Ngan 2007; Liu et al. 2007) and network bandwidth allocation (Chong et al. 1995; Adas 1998; Wu et al. 2001)). In Liu et al. (2007), the authors studied relationship between the peak signal-to-noise ratio and quantization parameter and propose a linear rate-quantization model to optimize quantization parameter calculation. In Wu et al. (2001), the authors presented a dynamic network resource allocation scheme for high-quality variable bitrate video transmission, based on the prediction of future traffic patterns. While monitoring and controlling QoS parameters of the video transmission system is important for achieving high video quality, it is more crucial to evaluate video quality from the users' perspective, which is known as QoE, or user-lever QoS. QoE-based video quality assessment for wireless video streaming is difficult because user experience is subjective and hard to quantify and measure. Moreover, the advent of new video compression standards, the development of video transmission systems, and the advancement of consumer video technologies all call for a newer and better understanding of user QoE. For the study of QoE, Li et al. (2018) proposed a subjective QoE metric for 360 VR video transmission over wireless networks. It makes a pixel-by-

pixel comparison between the reference content and the distorted content, producing a peak SNR figure without considering what the contents actually represent (Li et al. 2016).

In general, subjective QoE assessment tests are the most common method of studying video subjective QoE. The subjective QoE assessment method is simply that the video users watch the video in a controlled environment and use Mean Opinion Score (MOS) to score the video quality. ITU-T (Methodology for the Subjective Assessment 2002) has proposed several provisions for assessment protocol and experimental environment for subjective video quality assessment, and MOS score (Tan et al. 2016) is considered to be the most accurate reflection of video quality perception through the human visual system (HVS). The protocols for subjective video quality assessment can be generally classified into two categories: the methods of using Single Stimulus (SS) and Double Stimulus (DS). Single Stimulus (SS) method can be used when only processed videos presented to the subject. Single Stimulus Continuous Quality Scale (SSCQS) (Seshadri-nathan et al. 2010) is one of the representative Single Stimulus (SS) methods.

Unlike subjective QoE models, the study of objective QoE models does not require evaluative experiments; the relevant mathematical metrics only need to be studied. For 2D video, PSNR (or SSIM) is the most used objective QoE evaluation metric, which is based on the mean squared errors (MSE) between the reference and processed videos. However, PSNR cannot successfully reflect the subjective visual quality perceived by the human visual system (HVS) because it does not consider human visual perception at all. In order to correlate objectivity assessment with subjective quality better, many improved methods (Xu et al. 2014; Liu and Zhai 2017) based on PSNR have been proposed. Xu et al. (2014) proposed a weight-based PSNR metric to measure the quality of video. When calculating PSNR, their method imposes greater penalty weight on areas with faces and facial features. The free energy adjusted PSNR (FEA-PSNR) is proposed in Liu and Zhai (2017), which considers image perceptual complexity when evaluating

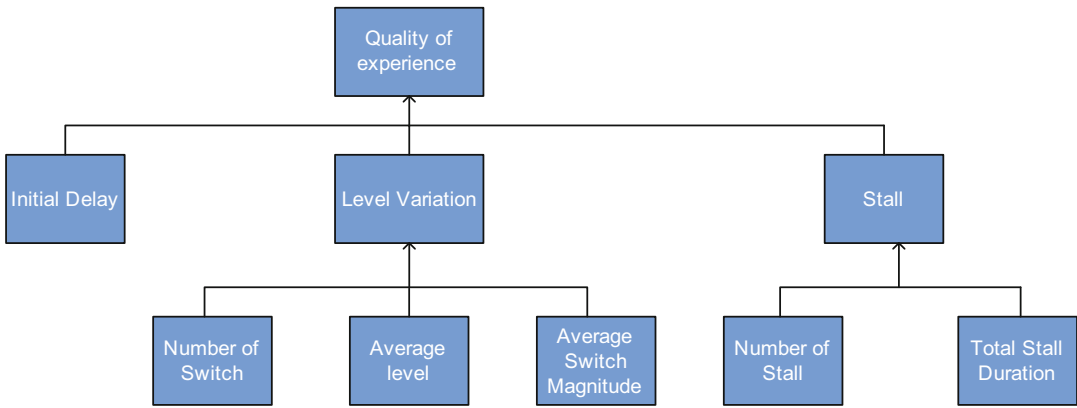
image quality. Furthermore, Na and Kim (2014) proposed a non-reference PSNR method to evaluate quantization error and the blocky effect when measuring nonreference quality of H.264 / AVC videos. Lee et al. (2016) explore the applicability of some existing objective models developed for 2D video to perceptual 3D video quality measurement.

Foundations

Researchers have studied the QoE metric for traditional single view video streaming and proposed different QoE metrics. In Oyman and Singh (2012), the recently standardized QoE metrics and reporting framework in third Generation Partnership Project (3GPP) were reviewed. The 3GPP dynamic adaptive streaming over HTTP (DASH) specification provides mechanisms for QoE measurements at the client device as well as protocols and formats for delivery of QoE reports to the network servers. Piamrat et al. (2009) focused on a hybrid QoE assessment model that retained the advantages of both subjective and objective schemes but minimized their drawbacks based on statistical learning using a random neural network (RNN).

In general, the QoE influence factors can be categorized into technical and perceptual influence factors. The perceptual influence factors are directly perceived by the end user of the application and are dependent but decoupled from the technical development. For example, several technical reasons can introduce initial delay, but the end user only perceives the waiting time. Therefore, it is necessary to analyze also the technical influence factors, which drive the perception of the end users. In general, QoE can be measured either by objective monitoring or subjective test. In video transmission system, QoE is influenced by many factors such as video capture, encoding, delivery, decoding, rendering, display, and context of use.

We give a brief introduce of the influencing factors of QoE for wireless video streaming. As shown in Fig. 1, we investigate the three major of influencing factors: initial delay, level variation, and stall.



Quality of Experience for Wireless Video Streaming, Fig. 1 The influencing factors of QoE for wireless video streaming

- *Initial Delay:* During the wireless video streaming process, there is an initial delay before the first frame of the video can be displayed, due to the need for the video client to buffer a certain amount of video data (Liu et al. 2016).
- *Level Variation:* During a video session, the video quality might keep switching. We propose three dimensions for factor level variation: (1) average level, which indicates the average quality of the entire video session; (2) number of switches, which indicates the frequency of quality switch; (3) average switch magnitude, which indicates the average amplitude of quality change.
- *Stall:* During a video session, it is possible that there is the network bandwidth fluctuation, leading to buffer underflow and stalling (rebuffering). As to the stall, most of the current research studies only consider the relationship between stall duration and QoE. However, it is obvious that the number of stalls is also vital.

Finally, QoE can be measured either by objective monitoring or subjective test. The subjective test is to obtain the quantitative representation of QoE by asking video users to grade the received video contents during the video playback in a controlled environment. It is believed to be the most accurate reflection of video quality perceived by users.

Subjective test directly measures user QoE by soliciting users' evaluation scores under the laboratory environment. Users are given a series of tested video sequences, original ones, and processed ones included and then required to give scores on the video quality. Detailed plans for conducting subjective tests have been made by the Video Quality Expert Group (VQEG) (1998). Though being viewed as a relative accurate way of measuring user QoE, subjective test suffers from three major drawbacks. First, subjective test has high cost in terms of time, money, and manual effort. Second, subjective test is conducted in the laboratory environment, with limited test video types, test conditions, and viewer demography. Therefore, the results may not be applicable to video quality assessment in the wild. Third, the subjective test cannot be used for real-time QoE evaluation.

In order to avoid high cost of subjective test, objective quality models have been developed. The major purpose is to identify the objective QoE parameters that contribute to user perceptual quality. Subjective test results are often used as ground truth to validate the performance of the objective quality models. Most of the objective quality models are based on how the Human Visual System (HVS) receives and processes the information of the video signals. One of the commonly

used methods is to quantify the difference between the original video and the distorted video, then weight the errors according to spatial and temporal features of the video. However, the need to access original video hinders online QoE monitoring. In order to develop QoE prediction models that do not depend on original videos, network statistics (such as packet loss) and spatiotemporal features extracted or estimated from the distorted video are leveraged. In the objective approach, mathematical models (algorithms) based on the human visual system are created to assess the quality of videos without external interferences. In image and video processing, PSNR is the most used objective quality assessment metric. It makes a pixel-by-pixel comparison between the reference content and the distorted content, producing a peak SNR figure without considering what the contents actually represent.

Key Application

Quality of Experience for wireless video streaming can optimize video transmission. In QoE for wireless video streaming, one of the most important aspects of studying QoE is how to improve or maximize users' overall QoE value. How to improve the experience of user is also the significance of our research of quality of experience for wireless video streaming.

Future Directions

There are some aspects that need us to research about QoE for wireless video streaming in the future. The biggest problem is that the end user's perceived instantaneous (or overall) QoE is not being objectively measured (or maximized). Although reducing the number of stalls and bitrate switches is a reasonable approach to reduce viewer annoyance, it does not capture a viewer's holistic perception of quality, nor does it guarantee the best (instantaneous or overall) QoE. A user's perceived QoE is greatly

influenced by a complex interplay of video content, varying numbers of rebuffering events, highly diverse rebuffering lengths, the rebuffering locations within a video, and so on. Therefore, having a fast and accurate objective QoE predictor that automatically predicts instantaneous quality scores that correlate well with perceived time-varying QoE could serve as feedback information that could be used to control and improve the performance of methods of quality of experience for wireless video streaming.

Cross-References

► Quality of Service

References

- Adas AM (1998) Using adaptive linear prediction to support real-time VBR video under RCBP network service model. *IEEE/ACM Trans Networking* 6(5): 635–644
- Chen Z, Ngan KN (2007) “Recent advances in rate control for vide coding,” *signal process. Image Commun* 22(1):19–38
- Chong S, Li S-Q, Ghosh J (1995) Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM. *IEEE J Sel Areas Commun* 13(1):12–23
- Jin J, Nahrstedt K (2004) Qos specification languages for distributed multimedia applications: a survey and taxonomy. *IEEE MultiMedia* 11(3):74–87
- Lee C, Ok J, Youn S, Seo G (2016) Objective perceptual quality assesment and depth perception of 3D media. In: 2016 international conference on telecommunications and multimedia (TEMU), Heraklion, pp 1–5. <https://doi.org/10.1109/TEMU.2016.7551929>
- Li J, Bao Z, Zhang C, Li Q, Liu Z (2016) Joint mcs and power allocation for svc video multicast over heterogeneous cellular networks. *Comput Commun* 83(C): 16–26
- Li J, Feng R, Liu Z, Sun W, Li Q Modeling QoE of virtual reality video transmission over wireless networks, accepted by *IEEE Global Communications Conference*, July 2018
- Liu N, Zhai G (2017) Free energy adjusted peak signal to noise ratio (FEA-PSNR) for image quality assessment. *Sens Imaging* 18(1):11
- Liu Y, Li ZG, Soh YC (2007) A novel rate control scheme for low delay video communication of H.264/AVC

standard. IEEE Trans Circuits Syst Video Technol 17(1):68–78

Liu Z, Dong M, Gu B, Zhang C, Ji Y, Tanaka Y (2016) Fast-start video delivery in future internet architectures with intra-domain caching. Mob Netw Appl 22(1): 98–112

ITU-R Recommendation BT.500-7: Methodology for the subjective assessment of the quality of television pictures, ITU, Geneva, Switzerland, 1995

Na T, Kim M (2014) A novel no-reference PSNR estimation method with regard to deblocking filtering effect in H. 264/AVC bitstreams. IEEE Trans Circuits Syst Video Technol 24(2):320–330

Oyman O, Singh S (2012) Quality of experience for http adaptive streaming services. IEEE Commun Mag 50(4):20–27

Piamrat K, Viho C, Bonnin JM, Ksentini A (2009) Quality of experience measurements for video streaming over wireless networks. In: 2009 sixth international conference on information technology: New Generations, pp 1184–1189

Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) A subjective study to evaluate video quality assessment algorithms. In: IS&T/SPIE Electronic Imaging, pp 75 270H–75 270H

Sullivan GJ, Wiegand T (1998) Rate-distortion optimization for video compression. IEEE Signal Process Mag 15(6):74–90

Tan TK, Weerakkody R, Mrak M, Ramzan N, Baroncini V, Ohm J-R, Sullivan GJ (2016) Video quality evaluation methodology and verification testing of hevc compression performance. IEEE Trans Circuits Syst Video Technol 26(1):76–90

VQEG, Final report from the video quality experts group on the validation of objective models of video quality assessment, March 2000. <http://www.vqeg.org/>

Wu M, Joyce RA, Wong H-S, Guan L, Kung S-Y (2001) Dynamic resource allocation via video content and short-term traffic statistics. IEEE Trans Multimedia 3(2):186–199

Xu M, Deng X, Li S, Wang Z (2014) Region-of-interest based conversational HEVC coding with hierarchical perception model of face. IEEE J Sel Top Signal Processing 8(3):475–489

Zhao T, Liu Q, Chen CW (2017) Qoe in video transmission: a user experience-driven strategy. IEEE Commun Surveys Tutor 19(1):285–302

Quality of Protection

► Long-Term Quality of Protection

Quality of Service

► Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes

Quality of Service (QoS) for Wireless Video Streaming

► Quality of Experience for Wireless Video Streaming

Quality of Service in IEEE 802.11 Networks

Xianghui Cao
Southeast University, Nanjing, China

Synonyms

QoS enhancements in IEEE 802.11e; QoS provisioning in IEEE 802.11

Definitions

In order to provision quality of service (QoS) in IEEE 802.11 networks, several enhancements to the original IEEE 802.11 standard have been proposed to provide QoS services for multimedia and real-time applications over WLANs.

Historical Background

The IEEE 802.11 specification was first published in 1997 (Hiertz et al. 2010). After a

Quality of Experience (QoE) for Wireless Video Transmission

► Quality of Experience for Wireless Video Streaming

series of subsequent amendments, nowadays, it has become one of the most widely deployed wireless technologies. Traditional IEEE 802.11 standards are difficult to provide QoS services for users, due to the lack of an explicit mechanism for service differentiation. For example, the IEEE 802.11a release and also the following IEEE 802.11b and IEEE 802.11g all employ the distributed coordination function (DCF)-based schemes, in which access to the medium is given on a first-come-first-served (FCFS) basis. The FCFS mechanism could not distinguish the traffic flows with different QoS requirements; in fact, there is no notion of traffic priority, indicating no guarantee of QoS in IEEE 802.11a/b/g. This may result in performance degradation during heavy-load periods, especially in multimedia applications.

In 2005, IEEE 802.11e is released explicitly targeting at provisioning QoS solutions to IEEE 802.11 networks (IEEE 802.11e 2005). As an important extension to the traditional IEEE 802.11 standard, the IEEE 802.11e release incorporates several QoS enhancements in the media access control (MAC) layer. It improves the MAC layer for QoS provisioning in which data packets are prioritized based on their application requirements. In IEEE 802.11e, the DCF and the optional point coordination function (PCF) are combined and enhanced to a new version called hybrid coordination function (HCF).

In 2016, the most recent IEEE 802.11 release has incorporated HCF in the MAC layer (IEEE 802.11 2016).

Foundations

IEEE 802.11e enhances the MAC layer functionality by integrating HCF to provide prioritized and parameterized QoS access to the wireless medium specified by the IEEE 802.11 physical layer. HCF combines functions from the DCF and PCF in original 802.11 MAC protocols for both contention access and contention-free access. Specifically, it consists of two mechanisms for service differentiation,

namely, enhanced distributed channel access (EDCA) for contention-based channel access and HCF controlled channel access (HCCA) for contention-free traffic delivery (IEEE 802.11e 2005).

Enhanced Distributed Channel Access (EDCA)

The EDCA mechanism defines eight user priorities (values assigned for each data unit) and prioritizes traffic into four access categories (ACs), as listed in Table 1.

With EDCA, stations belonging to different ACs can contend for sending packets with different sets of parameters, including an arbitrary interframe space (AIFS), CW_{min} and CW_{max} (i.e., minimum and maximum contention window limits, respectively), and transmission opportunity (TXOP) limit. The default parameters specified by the IEEE 802.11e standard are listed in Table 2. AIFS determines the amount of time a station senses the channel to be idle before conducting backoff or transmitting packets, the contention window sizes define the bounds of random backoff periods, and TXOP limit defines the duration a station can transmit after acquiring the channel.

- *AIFS*: IEEE 802.11 employs a DCF interframe space (DIFS) for which each frame must wait until the medium is available via Clear Channel Assessment (CCA). Instead of using a constant DIFS, EDCA assigns a shorter IFS to a station with a higher priority so that it can

Quality of Service in IEEE 802.11 Networks, Table 1
Mapping between priorities and ACs

User priority	Access category (AC)	Designation
1 (lowest)	AC_BK	Background
2	AC_BK	Background
0	AC_BE	Best effort
3	AC_BE	Best effort
4	AC_VI	Video
5	AC_VI	Video
6	AC_VO	Voice
7 (highest)	AC_VO	Voice

Quality of Service in IEEE 802.11

Networks, Table 2

Default parameters in EDCA

Access category (AC)	CWmin	CWmax	AIFSN
AC_BK	aCWmin	aCWmax	7
AC_BE	aCWmin	aCWmax	3
AC_VI	$(aCWmin + 1)/2 - 1$	aCWmin	2
AC_VO	$(aCWmin + 1)/4 - 1$	$(aCWmin + 1)/2 - 1$	2

wait for less time before sending packets, implying a shorter expected delay than others with lower priorities. AIFS is decided by the access category and an arbitration interframe spacing number (AIFSN) as $AIFS[AC] = AIFSN[AC] \times aSlotTime + aSIFSTime$, where the parameter AIFSN is assigned by the AP or a management entity.

- *Contention window limits:* In EDCA, the contention window limits can be assigned distinct for different ACs. As shown in Table 2, with the default settings, ACs for heavy traffic in video and voice applications can be assigned smaller contention windows to let the corresponding stations wait less time (on average) in conducting backoff.
- *TXOP limit:* A TXOP is gained after a station succeeds in the channel access contention. It is defined by a starting time and a maximum duration, which controls the time that the station can uninterruptedly occupy the channel and transmit packets. Differentiation in the TXOP parameter can allow heavy-traffic stations to transmit multiple packets for a longer time and prevent low-load stations to occupy the channel for an unnecessarily period of time.

HCF Controlled Channel Access (HCCA)

The HCCA mechanism is developed from the point coordination function (PCF) in traditional IEEE 802.11. It makes use of a high-priority hybrid coordinator (HC), a centralized controller that collocates with the AP, to transmit packets to stations and control the channel access of them by allocating TXOPs. In order to support QoS, HCCA allows stations to request the HC for TXOPs to transmit packets to or download packets from the AP based on their own QoS demands. The HC decides whether to accept a request based on an admission control

mechanism, to be discussed below, and schedules TXOPs for the corresponding station if its request is accepted. Unlike PCF that divides a beacon interval into a control-free period (CFP) and a control period (CP), HCCA defines an additional controlled access phase (CAP) that allows the HC to reserve the channel for contention-free packet exchanges with stations that have been granted TXOPs during both CFP and CP. After the HC senses the channel to be idle for a PCF interframe space (PIFS), it transmits a packet with the duration value set to cover the CAP to claim the channel access. PIFS is smaller than DIFS and AIFS, giving the HC priority over other traffics. Therefore, a CAP may be initiated at any time, and there can be multiple CAPs within a CP. Other stations in the CP apply normal EDCA to access the channel and transmit packets.

Admission Control

The service differentiation at MAC layer is effective for QoS provisioning when network load is moderate (Malik et al. 2015). However, in situations with heavy multimedia traffic, pure service differentiation becomes inefficient and may incur long congestion delay (Zhu et al. 2004). In this case, admission control is necessary, which selectively admits traffic flows such that the QoS performance of existing flows is still ensured. There are two admission control mechanisms specified in IEEE 802.11e for contention-based access and controlled access, respectively. In the former mechanism, stations in ACs that require admission control send requests to the HC for admitting their traffic flows, while in the latter mechanism, stations request the HC for TXOPs as mentioned above (IEEE 802.11e 2005). Precise admission control is usually difficult since the accurate network status such as channel conditions, throughput, and delay may be not available. To overcome this difficulty, several methods have been pro-

posed, including measurement-based admission control based on continuously monitoring network status and model-based admission control based on some analytical or empirical models (Gao et al. 2005).

Related Work

The IEEE 802.11 standards focus on the physical and MAC layers and provide QoS support mainly in the MAC layer. On top of these layers, many QoS issues and mechanisms have been proposed such as QoS routing, congestion control, rate control, and energy-efficient QoS solutions (Peng et al. 2013; Shen et al. 2013; El Masri et al. 2014; Luo and Shyu 2011). Recently, QoS solutions can be provided within the software-defined networking (SDN) and cloud computing frameworks. For example, researchers have proposed a QoS solution called QoS Flow for OpenFlow-based SDN networks through the appropriate control of packet scheduling. In addition, OpenQoS (Egilmez et al. 2012) is an OpenFlow controller designed for supporting multimedia flows with end-to-end QoS requirements, which presents a new dynamic QoS routing scheme that maintains the shortest path for the data delivery to minimize packet loss and latency. In cloud-based wireless networks, a precise modeling of the genuine surroundings of IEEE 802.11 WLANs is important for proficient QoS-aware cloud service provisioning (Tursunova and Kim 2012). However, since these are general issues for all networks, they are not explicitly covered in this chapter.

Key Applications

QoS guarantee is a critical requirement for throughput-sensitive and delay-sensitive applications such as multimedia and real-time factory communications over WLANs.

Cross-References

► [QoS-Aware MAC](#)

References

- Egilmez HE, Dane ST, Bagci KT, Tekalp AM (2012) OpenQoS: an openflow controller design for multimedia delivery with end-to-end quality of service over software-defined networks. In: Signal & information processing association annual summit and conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, pp 1–8
- El Masri A, Sardouk A, Khoukhi L, Hafid A, Gaiti D (2014) Neighborhood-aware and overhead-free congestion control for IEEE 802.11 wireless mesh networks. *IEEE Trans Wirel Commun* 13(10):5878–5892
- Gao D, Cai J, Ngan KN (2005) Admission control in IEEE 802.11e wireless LANs. *IEEE Netw* 19(4):6–13
- Hiertz GR, Denteneer D, Stibor L, Zang Y, Costa XP, Walke B (2010) The IEEE 802.11 universe. *IEEE Commun Mag* 48(1):62–70
- IEEE 80211 (2016) IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012) - IEEE standard for information technology-telecommunications and information exchange between systems Local and metropolitan area networks-specific requirements - part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications.
- IEEE 80211e (2005) ANSI/IEEE 802.11e-2005 - IEEE standard for information technology-local and metropolitan area networks-specific requirements-part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications - amendment 8: medium access control (MAC) quality of service enhancements
- Luo H, Shyu ML (2011) Quality of service provision in mobile multimedia—a survey. *Hum Centric Comput Inf Sci* 1(1):5
- Malik A, Qadir J, Ahmad B, Yau KLA, Ullah U (2015) QoS in IEEE 802.11-based wireless networks: a contemporary review. *J Netw Comput Appl* 55:24–46
- Peng Y, Yu Y, Guo L, Jiang D, Gai Q (2013) An efficient joint channel assignment and QoS routing protocol for IEEE 802.11 multi-radio multi-channel wireless mesh networks. *J Netw Comput Appl* 36(2):843–857
- Shen X, Cheng X, Zhang R, Jiao B, Yang Y (2013) Distributed congestion control approaches for the IEEE 802.11p vehicular networks. *IEEE Intell Transp Syst Mag* 5(4):50–61
- Tursunova S, Kim YT (2012) Realistic IEEE 802.11e EDCA model for QoS-aware mobile cloud service provisioning. *IEEE Trans Consum Electron* 58(1)
- Zhu H, Li M, Chlamtac I, Prabhakaran B (2004) A survey of quality of service in IEEE 802.11 networks. *IEEE Wirel Commun* 11(4):6–14

Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes

Han-Bae Kong¹, Ping Wang^{2,3}, and Dusit Niyato⁴

¹School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

²School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore, Singapore

³Department of Information and Communication Engineering, Tongji University, Shanghai, China

⁴School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore, Singapore

Synonyms

Ginibre point process; Quality of service; Repulsive point process; Stochastic geometry; Wireless sensor network

Definitions

Quality of service (QoS) indicates the description of the performance of a service which is influenced by various features such as throughput, errors, and latency. Wireless sensor network (WSN) is a network consisting of spatially distributed sensor nodes which monitor and report physical or environmental conditions. Ginibre point process (GPP) is a point process which can consider repulsion among the locations of the points in the point process.

Historical Background

Mathematically analyzing QoS of wireless networks helps us to obtain insights on the impacts of systems parameters on the QoS. Recently, stochastic geometry has been widely used as a tool for modeling wireless networks and analyzing QoS of the networks (Haenggi 2012). Due to its simplicity, the Poisson point process (PPP), whose the locations of the points are independent,

has been applied to various wireless networks such as ad hoc wireless networks (Weber et al. 2007), cellular networks (Dhillon et al. 2012), and wireless energy harvesting networks (Sakr and Hossain 2015).

In practical networks, the locations of transmitters may not be placed close to each other in order to mitigate interference and extend coverage region (Lin et al. 2015). In this context, determinantal point processes (DPPs) (Li et al. 2015) which can consider this repulsive nature have gained a lot of attention as a model for wireless networks. The GPP is one example of the DPP (Decreusefond et al. 2015). The α -GPP ($-1 \leq \alpha < 0$) is a DPP which contains the GPP ($\alpha = -1$) and the PPP ($\alpha \rightarrow 0$) as special cases. In the α -GPP, the parameter α (termed *repulsion parameter*) represents the degree of the repulsion among the points, and the strength of the repulsion grows as α decreases. In Kong et al. (2016), the performance of ambient wireless energy harvesting networks was investigated by adopting the α -GPP.

QoS Analysis

Fundamental Properties of the α -GPP

$B(r)$ denotes a ball centered at the origin with radius r . Then, for an α -GPP Φ on an observation window $W = B(R)$ with intensity λ , the hole probability is given by

$$Pr(\Phi \cap B(r) = \emptyset) = Pr(\forall x \in \Phi, \|x\| > r) \\ = \prod_{n \geq 0} \left(1 + \alpha \frac{\gamma(n+1, \pi \lambda r^2)}{n!} \right)^{-1/\alpha}, \quad (1)$$

where $\gamma(a, b) = \int_0^b t^{a-1} e^{-t} dt$ is the lower incomplete gamma function. Also, for a non-negative function q , the Laplace transform is expressed as

$$E \left[\exp \left(- \sum_{x \in \Phi} q(x) \right) \right] = \prod_{n \geq 0} (1 + \alpha v_n)^{-1/\alpha}, \quad (2)$$

where

$$v_n = \frac{2(\pi\lambda)^{n+1}}{n!} \int_0^R (1 - \exp(-q(x))) \exp(-\pi\lambda r^2) r^{2n+1} dr.$$

System Model

WSNs are composed of a number of gateways and sensor nodes. It is assumed that each sensor node is connected to the closest gateway. In order to consider a repulsive nature in the WSNs, the spatial distributions of the gateways and the sensor nodes are modeled by independent α -GPPs Φ_g and Φ_s with intensities λ_g and λ_s and repulsion parameters α_g and α_s , respectively.

On a given resource block, a single sensor node is associated with the nearest gateway, and therefore for a sensor node located at y , the association rule can be expressed as

$$\begin{aligned} x^o &= \arg \min_{x \in \Phi_g} \|x - y\|, \\ d &= \min_{x \in \Phi_g} \|x - y\|, \end{aligned} \quad (3)$$

where x^o denotes the location of the associated gateway. In order to alleviate interference, sensor nodes employ a fractional channel inversion power control. More specifically, the transmit power at a sensor node at y is

$$P = \tau d^\epsilon, \quad (4)$$

where τ and $\epsilon \in [0, 1]$ indicate the receiver sensitivity and the power control factor, respectively.

The throughput of a link is related to the signal-to-interference-plus-noise ratio (*SINR*) of the link. Therefore, the *SINR* outage probability is one of the most widely used QoS metrics.

From the stationarity of the α -GPP, without loss of generality, it is assumed that the gateway serving a typical sensor node (termed *tagged gateway*) is located at the origin. Then, the *SINR* outage probability at the tagged gateway is given by

$$P_{\text{out}} = \Pr(\text{SINR} < \gamma_{th}), \quad (5)$$

where

$$\text{SINR} = \frac{\tau h_0 \|x_0\|^{l(\epsilon-1)}}{\sum_{x \in \Phi_I} P_x h_x \|x\|^{-l} + \sigma^2}. \quad (6)$$

Here, h_0 , l , and σ^2 denote the power of the fading channel between the tagged gateway and the typical sensor node, the path loss exponent, and the power of the additive white Gaussian noise, respectively. Φ_I is a point process which presents the spatial distribution of the interfering sensor nodes. Also, x_0 , P_x , and h_x are the location of the typical sensor node, the transmit power at the sensor node at x , and the power of the fading channel between the tagged gateway and interfering sensor at x . It is assumed that fading channels follow the Rayleigh distribution, and hence h_0 and $\{h_x\}$ follow i.i.d. exponential distribution with unit mean.

Outage Probability

As seen in (5), the *SINR* outage probability is related to both the distances between sensor nodes and their serving gateways and the transmit powers at sensor nodes.

Theorem 1 *The cumulative distribution function (CDF) and the probability density function (PDF) of d in (3) are, respectively, given by*

$$F_d(r) = \Pr(d \leq r) = 1 - \prod_{n \geq 0} \left(1 + \alpha_g \frac{\gamma(n+1, \pi\lambda_g r^2)}{n!} \right)^{-1/\alpha_g}, \quad (7)$$

$$\begin{aligned} f_d(r) &= 2 \exp(-\pi\lambda_g r^2) (1 - F_d(r)) \\ &\quad \times \sum_{n \geq 0} \frac{1}{n!} (\pi\lambda_g)^{n+1} r^{2n+1} \left(1 + \alpha_g \frac{\gamma(n+1, \pi\lambda_g r^2)}{n!} \right)^{-1}. \end{aligned} \quad (8)$$

Proof. The CDF in (7) can be derived from (1), and the PDF in (8) can be obtained by differentiating the CDF in (7) with respect to r . $\square$

Theorem 2 The CDF and the PDF of P in (4) are, respectively, given by

$$F_P(p) = 1 - \prod_{n \geq 0} \left(1 + \alpha_g \frac{\gamma(n+1, \pi \lambda_g p^{\frac{2}{\varepsilon l}} \tau^{-\frac{2}{\varepsilon l}})}{n!} \right)^{-1/\alpha_g}, \quad (9)$$

$$f_P(p) = \frac{2}{\varepsilon l} \exp\left(-\pi \lambda_g p^{\frac{2}{\varepsilon l}} \tau^{-\frac{2}{\varepsilon l}}\right) (1 - F_P(p)) \sum_{n \geq 0} \frac{1}{n!} \left(\pi \lambda_g \tau^{-\frac{2}{\varepsilon l}}\right)^{n+1} \\ \times p^{\frac{2}{\varepsilon l}(n+1)-1} \left(1 + \alpha_g \frac{\gamma\left(n+1, \pi \lambda_g p^{\frac{2}{\varepsilon l}} \tau^{-\frac{2}{\varepsilon l}}\right)}{n!} \right)^{-1}.$$

Proof. The CDF in (9) can be identified from (1) and (4). In addition, the PDF in (10) can be derived by differentiating the CDF in (9) with respect to p . $\square$

From the CDFs in (7) and (9), it is shown that the CDFs are decreasing functions of α_g since $\alpha_g < 0$. Therefore, for a given sensor node, the distance between the sensor node and its serving gateway and the transmit power at the sensor node get higher as the degree of the repulsion among gateways decays (i.e., α_g increases). In this context, from the expression of the SINR in (6), it is expected that the SINR outage probability P_{out} is an increasing function

of α_g since an increase in α_g leads to an increase of $\|x_0\|$ and P_x .

In order to obtain an exact expression of P_{out} in (5), the distribution of Φ_I should be identified. However, the distribution is still unknown. Since a single sensor node is associated with a gateway at a given resource block, the intensity of Φ_I is equal to λ_g . Motivated by the fact, Φ_I can be approximated by an α -GPP with intensity λ_g and repulsion parameter $\alpha_s \tilde{\Phi}_I$. Then, an approximation of P_{out} in (5) can be derived as presented in the following theorem.

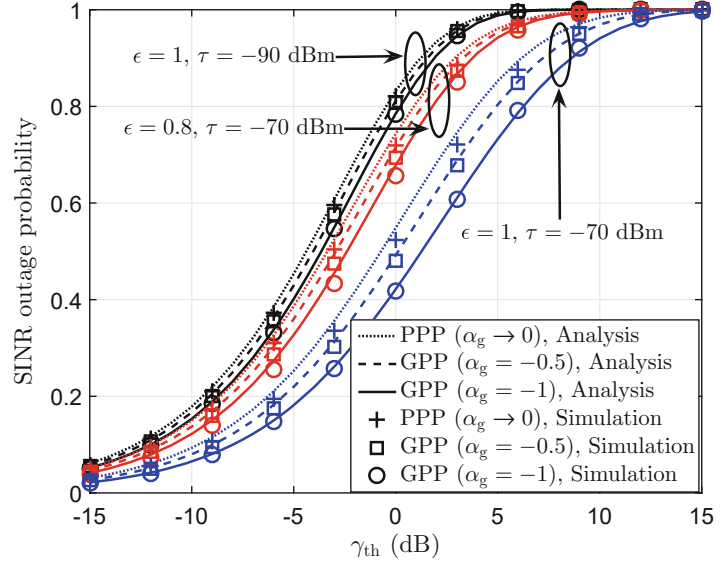
Theorem 3 An approximation of the SINR outage probability P_{out} in (5) is given by

$$P_{\text{out}} \approx 1 - \int_0^\infty \exp\left(-\frac{\gamma_{th} \sigma^2}{\tau u^{l(\varepsilon-1)}}\right) K(u) f_d(u) du, \quad (11)$$

$$K(u) = \prod_{n \geq 0} \left(1 + \frac{2\alpha_s (\pi \lambda_g)^{n+1}}{n!} \right. \\ \left. \times \int_0^\infty \int_{\left(\frac{p}{\tau u^{l(\varepsilon-1)}}\right)^{\frac{1}{l}}}^R \exp\left(-\pi \lambda_g r^2\right) \frac{1}{1 + \tau r^l u^{l(\varepsilon-1)}/(\gamma_{th} p)} r^{2n+1} dr f_P(p) dp \right)^{-1/\alpha_s}.$$

Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes,

Fig. 1 SINR outage probability as a function of γ_{th} when Φ_s is a PPP ($\alpha_s \rightarrow 0$), $\lambda_g = 4$ gateways/km², $\sigma^2 = -90$ dBm, $R = 5$ km, and $l = 4$



Proof. Since h_0 is an exponential random variable with unit mean,

$$P_{\text{out}} = \Pr \left(h_0 < \frac{\gamma_{th}}{\tau \|x_0\|^{l(\epsilon-1)}} (I + \sigma^2) \right) \\ = 1 - \int_0^\infty \exp \left(-\frac{\gamma_{th} \sigma^2}{\tau u^{l(\epsilon-1)}} \right) L_I \left(\frac{\gamma_{th}}{\tau u^{l(\epsilon-1)}} \right) f_d(u) du, \quad (12)$$

where $I = \sum_{x \in \Phi_I} P_x h_x \|x\|^{-l}$ and $L_X(s) = E[e^{-sX}]$ denote the Laplace transform of a random variable X . Then, by approximating the distribution of interferers as $\tilde{\Phi}_I$ and by applying the formula in (2), an approximation of P_{out} can be obtained as in (11). For the details, please refer to Kong et al. (2017). $\square$

From Fig. 1, it is observed that the approximation in (11) shows almost identical performance with the simulated results. Note that a higher α_g results in higher transmit powers at sensor nodes and P_{out} is an increasing function of α_g . Therefore, it is concluded that QoS of WSNs can be improved with lower transmit powers when there exists repulsion among the locations of the gateways.

Cross-References

- [Long-Term Quality of Protection](#)
- [QoE Assessment Models](#)
- [QoS-Aware MAC](#)
- [QoS in Indoor Localization](#)
- [Quality of Service in IEEE 802.11 Networks](#)

References

- Decreusefond L, Flint I, Vergne A (2015) Efficient simulation of the Ginibre point process. *Adv Appl Probab* 52:1–21
- Dhillon H, Ganti R, Baccelli F, Andrews J (2012) Modeling and analysis of K -tier downlink heterogeneous cellular networks. *IEEE J Sel Areas Commun* 30: 550–560
- Haenggi M (2012) *Stochastic geometry for wireless networks*. Cambridge University Press, Cambridge
- Kong H, Flint I, Wang P, Niyato D, Privault N (2016) Exact performance analysis of ambient RF energy harvesting wireless sensor networks with Ginibre point process. *IEEE J Sel Areas Commun* 34:3769–3784
- Kong H, Wang P, Niyato D, Cheng Y (2017) Modeling and analysis of wireless sensor networks with/without energy harvesting using Ginibre point processes. *IEEE Trans Wirel Commun* 16:3700–3713
- Li Y, Baccelli F, Dhillon H, Andrews J (2015) Statistical modeling and probabilistic analysis of cellular net-

works with determinantal point processes. *IEEE Trans Commun* 63:3405–3422

Lin T, Santoso H, Wu K (2015) Global sensor deployment and local coverage-aware recovery schemes for smart environments. *IEEE Trans Mobile Comput* 14:1382–1396

Sakr A, Hossain E (2015) Analysis of K -tier uplink cellular networks with ambient RF energy harvesting. *IEEE J Sel Areas Commun* 33:2226–2238

Weber S, Andrews J, Jindal N (2007) The effect of fading, channel inversion, and threshold scheduling on ad hoc networks. *IEEE Trans Inf Theory* 53:4127–4149

Quality-of-Service (QoS)

- [Effective Utilization of Unlicensed Spectrum](#)

Quantization

- [Wireless Key Establishment](#)

R

Radio Access Selection

- [Network Selection in Heterogeneous Wireless Networks](#)

Radio-Frequency Fingerprinting-Based Physical Layer Identification

Zhi Sun

Department of Electrical Engineering,
University at Buffalo, The State University of
New York, Buffalo, NY, USA

Synonyms

[User authentication](#); [User device identification](#)

Definitions

Physical layer identification (PLI) is the user device identification strategy based on the physical features of the wireless device. Radio-frequency fingerprinting is one of the most important PLI techniques that utilize the uniqueness in the waveform of the wireless signals emitted by the device to be identified.

Background

The wireless access to restricted networks, such as in enterprise or government, can grant great convenience and efficiency to the employees and visitors. However, the broadcasting nature of wireless communications brings significant secure risks, especially in the networks with sensitive data. Without the binding of wired connections, it is much easier for adversary to obtain security credentials from legitimate users through the wireless channel. Therefore, it is important to authenticate users not only by what they hold (e.g., logical level ID or pre-shared key) but also by what they are, i.e., identifying the wireless device itself based on various physical layer device features. Such device identification technique is defined as physical layer identification (PLI). PLI is a promising built-in secure access control solution that authenticates the device based on its inherent features. PLI can guarantee only authenticated devices access the network and can enhance the cryptography-based security protocols to authenticate users themselves.

PLI is not a single technique but a wide variety of user device identification techniques based on different physical features, including (i) the image artifacts in a photos taken by the camera of the user device, (ii) the response of the audio circuit to a predefined stimulus (e.g., a standard tone), (iii) the RF emissions from LCD screens, (iv) the measurement distortion caused by the MEMS in accelerometers or magnetome-

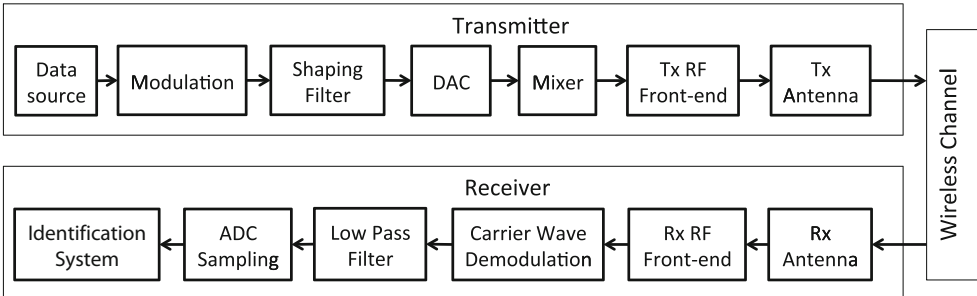
ters, and (v) the multipath effects added to signals when propagating through the wireless channel, among others. In this entry, we focus on one of the PLI techniques that we believe has the widest applicable range, i.e., the radio-frequency fingerprinting (RFF)-based PLI. RFF-based PLI identifies the user device based on any features extracted from the waveforms of the wireless signals, i.e., the electromagnetic waves that are radiated by the device to be tested. While the conventional software-level device identifications (e.g., IP or MAC address) can be easily changed by malwares, the RFF features cannot be modified without significant effort. Moreover, different from the other PLI techniques that either rely on the features from particular hardware not existing in any types of devices (e.g., camera, audio circuit, LCD, and accelerometers) or require the target device fixed in one location (e.g., multipath effects), the RFF features exist in any wireless device as long as it uses radio for communication and allow the user mobility. Therefore, the RFF-based technique is deemed as a promising wireless security solution.

In this entry, the sources of RFFs and the mechanisms to utilize such RFFs in the state-of-the-art RFF-based PLI techniques are explained through each step in the physical layer procedures in wireless communications.

Foundations

In essence, RFFs are random variables carrying discriminating information of user identity which

can be expressed in different feature domains, such as frequency domain, time domain, wavelet transform domain, etc. Figure 1 illustrates the logic procedure of RFF-based PLI during the wireless communication between a transmitter (the device to be identified) and a receiver (the identifier). The RFFs can come from many points along that procedure. At the transmitter side, the RFFs are rooted in the hardware imperfections of the transmitter device (Danev et al. 2012), which include clock jitter (Zanetti et al. 2010; Jana and Kasera 2010), the DAC sampling error (Polak et al. 2011), the mixer or local frequency synthesizer (Toonstra and Kinsner 1996, 1995), the power amplifier nonlinearity (Polak et al. 2011, Polak and Goeckel 2011; Liu and Doherty 2008), device antenna (Danev et al. 2009), and modulator sub-circuit (if the analog modulation is used) (Brik et al. 2008), among others. The EM waves stamped with the hardware-imperfection-based RFFs are then radiated by the transmitter antenna to the wireless channel, where the multipath effects of the wireless link can be also used as physical layer feature to identify user devices (Li et al. 2006; Xiao et al. 2007). However, the wireless channel-based physical layer feature only works when the user devices are fixed, while the RFF features support user mobility. At the receiver (i.e., the identifier), the RFF-stamped signals are received by the receiver antenna and processed through the analog circuits and digital processing units. Depending on their properties, different RFFs are extracted at (i) different parts of the signals, such as turn-on/off transient (Toonstra and Kinsner



Radio-Frequency Fingerprinting-Based Physical Layer Identification, Fig. 1 RFF-based PLI procedure along wireless communication flow block diagram

1995; Danev and Capkun 2009; Rehman et al. 2012a; Hall et al. 2004), data (Zanetti et al. 2010; Danev et al. 2009; Suski et al. 2008; Klein et al. 2009; Rehman et al. 2012b), and clock (Zanetti et al. 2010; Jana and Kasera 2010), and (ii) different domains, such as time (Brik et al. 2008), frequency (Danev et al. 2009; Danev and Capkun 2009), and wavelet domains (Brik et al. 2008). Finally, the extracted RFF is matched with the fingerprint database, which complete the physical layer identification process.

In the following sections, the complete RFF-based PLI procedure is systematically discussed in details. As shown in Fig. 1, the RFF-based PLI is closely coupled with wireless communications, including four major segments: the signal processing and RFF stamping at the transmitter, the signal propagation along wireless channels, the signal reception and feature extraction at the receiver, and the final identification and classification.

Signal Processing and RFF Stamping at Transmitter

At the beginning of the signal processing at the transmitter, the data source sends binary bits sequence to the digital modulation block, where the sequential binary bits are first mapped to higher-order and complex transmit symbols. We use m to index the sequence of the transmitted symbols so that each symbol is denoted as $x_m = x_m^I + j x_m^Q$, where x_m^I and x_m^Q are the in-phase and quadrature components, respectively. The duration of each symbol is determined by the expected data rate and the order of modulation scheme, which is denoted as T_{symbol} . Ideally, T_{symbol} should be a constant in one packet transmission. However, due to the clock-related hardware imperfection, the period/duration of each symbol can vary. As a result, the first possible RFF due to the hardware imperfection appears, which is regarded as a unique time domain feature and is defined as time interval error (TIE) (Zanetti et al. 2010). We denote this feature as σ_{TIE}^m at the m th symbol. Hence, the real symbol duration of the m th symbol becomes $T_m = T + \sigma_{TIE}^m$.

After that the complex symbols x_m is processed by the shaping filter $h_s(t, T_m)$, which gives the output $b_{\text{mod}}(x_m, h_s(t, T_m))$, where the formulation of the function $b_{\text{mod}}(\cdot)$ is determined by the specific modulation scheme. Then the digital signal moves to the digital-to-analog converter (DAC). It should be noted that $b_{\text{mod}}(x_m, h_s(t, T_m))$ is in fact a discrete function as the digital shaping filter $h_s(t)$ only changes value when the DAC generates new output. We consider that the DAC has a generation period T_g , i.e., the DAC converts the value of $b_{\text{mod}}(x_m, h_s(t, T_m))$ to analog current every T_g second. Then the input digital signal $u[n]$ (indexed by n) at the DAC can be expressed as

$$u[n] = A \sum_m b_{\text{mod}}(x_m, h_s(nT_g - mT_m, T_m)), \quad (1)$$

where A is the amplitude coefficient. It should be noted that the clock-related hardware imperfection could also influence the DAC generation period T_g . However, since T_g is much smaller than T_m (due to the high DAC sampling rate), the TIE in T_g is much less significant and can be neglected. It is also worth mentioning that any specific wireless protocol (e.g., 802.11, 802.15, GSM/GPRS, etc.) defines particular modulation schemes and shaping filter, which prepare the transmitted signal with different bandwidth and power spectrum density (PSD). Those parameters can dramatically influence the significance of RFF and eventually affect the performance of RFF-based PLI.

Ideally, the DAC converts the baseband signal sequence $u[n]$ to a time-continuous analog signal $u(t)$. However, the real output signal $y_u(t)$ from practical DAC is combed with quantization error and integral nonlinearity (INL) (D'Apuzzo et al. 2010), which can be expressed as

$$y_u(t) = \sum_{n=-\infty}^{\infty} (u(n) + \Delta_n) g\left(\frac{t - nT_g}{T_g}\right) + \Delta_{\text{INL}},$$

$$g(\theta) = \begin{cases} 1, & 0 \leq \theta < 1 \\ 0, & \text{elsewhere} \end{cases} \quad (2)$$

where Δ_n is quantization noise (for M-bit DAC quantization with input signal dynamic range $[-U, U]$, the maximum quantization error is $\delta_\Delta = 2^{-M} U$) and Δ_{INL} is the integral nonlinearity that is also regarded as a unique feature rooted in hardware imperfection in Polak et al. (2011).

After the DAC, the analog baseband signal is moved to passband by the mixer. Ideally, the complex passband signal $z(t)$ consists the in-phase component $i(t) \cdot \cos(\omega_c t)$ and the quadrature component $jq(t) \cdot \sin(\omega_c t)$, where ω_c is the carrier angle frequency and $i(t) = \text{Re}\{y_u(t)\}$, $q(t) = \text{Im}\{y_u(t)\}$. However, in practical, the phase offset quadrature error occurs in this procedure due to the imperfection of transmitter's frequency synthesizers (local oscillators), which modifies $z(t)$ to

$$\begin{aligned} z(t) = & \frac{1}{2} \left(\text{Re}\{y_u(t)\} \cdot e^{j\zeta/2} \right. \\ & \left. + \text{Im}\{y_u(t)\} \cdot e^{-j\zeta/2} \right) \cdot e^{j\omega_c t} \\ & + \frac{1}{2} \left(\text{Re}\{y_u(t)\} \cdot e^{j\zeta/2} \right. \\ & \left. - \text{Im}\{y_u(t)\} \cdot e^{-j\zeta/2} \right) \cdot e^{-j\omega_c t} \end{aligned} \quad (3)$$

where ζ is the quadrature error. In Toonstra and Kinsner (1995, 1996), the uniqueness of the mixer quadrature error is utilized as RFFs.

Finally, the passband signals go through the TX front-end amplifier and filter to gain enough power for radiation. In this procedure, another important RFF is stamped, i.e., the nonlinearity of the front-end amplifier (Gharaibeh 2011), which is widely used in many recent RFF-based PLI techniques (Gharaibeh 2011; Polak et al. 2011; Polak and Goeckel 2011; Liu and Doherty 2008). The output signal can be formulated as

$$\begin{aligned} w(t) &= h_{PA}(z(t), \tilde{\mathbf{a}}_{\text{tx}}) \otimes h_{BP}(t) \\ H_{BP}(f) &= \begin{cases} 1, & |f| < W_c \\ 0, & |f| > W_c \end{cases} \end{aligned} \quad (4)$$

where $h_{PA}(\cdot, \tilde{\mathbf{a}}_{\text{tx}})$ is nonlinear function of power amplifier at the transmitter, $\tilde{\mathbf{a}}_{\text{tx}}$ is complex power series, $\otimes$ stands for convolution computing, and

$h_{BP}(t)$ is the band-pass filter function that is nearly ideal within the carrier bandwidth W_c .

Signal Propagation Between TX and RX Antennas Along Wireless Channel

We consider the antennas as part of the wireless channel since both the antenna polarization and the multipath channel can be used as RFFs. With the input signal $w(t)$ given in (4), the intensity of the radiated EM waves by a polarized antenna is given by

$$\begin{aligned} \mathbf{A}_{tx}(t) &= F_{tx}^h[w(t)] \cdot e^{-j\phi_{tx}^h} \cdot \rho_h \\ &+ F_{tx}^v \cdot \rho_h + F_{tx}^v[w(t)] \cdot e^{-j\phi_{tx}^v} \cdot \rho_v \end{aligned} \quad (5)$$

where ρ_h and ρ_v are the direction unit vectors in the horizontal and vertical directions, respectively; $F_{tx}^h(\cdot)$ and $F_{tx}^v(\cdot)$ are the TX antenna projection functions that project the input signal $w(t)$ to the horizontal and vertical polarization, respectively; and ϕ_{tx}^h and ϕ_{tx}^v are the polarization phases at both directions. In some specific wireless system, such as RFID using the spiral coil antenna, the imperfection of antenna hardware can also contribute to the RFF (Danev et al. 2009). Moreover, even for the simpler antennas that are widely used in contemporary wireless devices (e.g., dipole antennas and patch antennas), although the RFF does not come from antenna imperfection, the random direction of antenna can generate random polarization. Then the antenna polarization can either influence the RFFs due to other hardwares (Danev and Capkun 2009) or act as a new RFF (coupled with the multipath effect) (Li et al. 2006; Xiao et al. 2007).

The radiated EM waves then propagate through the multipath wireless channel and finally reach the RX antenna at the receiver. As the RX antenna can be also polarized, the intensity of the EM waves at the RX antenna is also presented in both horizontal and vertical direction:

$$\mathbf{A}_{rx}(t) = \sum_{i=1}^{N_p} \mathbf{A}_{tx}(t) \cdot \left(h_i^h \cdot \rho_h + h_i^v \cdot \rho_v \right) \quad (6)$$

$$\otimes \delta(t - \tau_i)$$

where we consider there are N_p paths (including line-of-sight, reflected, diffracted, and refracted paths) connecting the transmitter and receiver; h_i^h and h_i^v are the path loss of the i th path with horizontal polarized waves and vertical polarized waves, respectively; and τ_i is the propagation delay of the i th path.

It is obvious that the complex multipath wireless channel can arbitrarily change the amplitude, the delay, and the phase of the EM waves at each single path. Consequently, the wireless channel can literally mask most, if not all, RFFs stamped at the transmitter. Hence, the channel effects on PLI need special attention in the performance evaluation. In Danev and Capkun (2009) and Danev et al. (2012), it has been proved that the PLI accuracy decreases as the distance between the transmitter and receiver increases. It should be noted that the channels considered in Danev and Capkun (2009) and Danev et al. (2012) are still friendly, which is open space without significant multipath and obstructions. The practical wireless channel, especially in real indoor or metropolitan environments, could be much more hostile in keeping the RFFs from the transmitter.

It should be also noted that another branch of PLI systems uses the multipath channel characteristics to authenticate the transmitter (mainly its location) (Li et al. 2006; Xiao et al. 2007). The obvious drawback is that such system only works when the transmitter and receiver are fixed and there should be no mobile obstructions in the environment that can change the channel characteristics. However, those conditions are not feasible in most current wireless applications.

Signal Reception and Processing at Receiver

At the receiver, the RX antenna captures the EM waves (given in (6)) and converts both the

horizontal and vertical polarized components to the received signal:

$$r(t) = F_{rx}^h{}^{-1} \left[\mathbf{A}_{rx}(t) \cdot \rho_h \cdot e^{j\phi_{rx}^h} \right] \quad (7)$$

$$+ F_{rx}^v{}^{-1} \left[\mathbf{A}_{rx}(t) \cdot \rho_v \cdot e^{j\phi_{rx}^v} \right];$$

where $F_{rx}^h(\cdot)^{-1}$ and $F_{rx}^v(\cdot)^{-1}$ are the RX antenna functions that capture the horizontal and vertical EM wave components, respectively, and ϕ_{rx}^h and ϕ_{rx}^v are the polarization phases of the RX antenna.

Similar as the transmitter, the received signal also needs to pass through the RF front-end at the receiver:

$$f(t) = h_{PA}(r(t) \otimes h_{BP}(t), \tilde{\mathbf{a}}_{rx}) \quad (8)$$

where $h_{PA}(\cdot, \tilde{\mathbf{a}}_{rx})$ is nonlinear function of power amplifier at the receiver and $h_{BP}(t)$ is the band-pass filter at the receiver, all similar to the case at the transmitter.

Then the output signal from the RX front-end is converted to baseband by the receiver mixer with quadrature errors (i.e., carrier demodulation). The demodulation signal then goes through the low-pass filter $h_{LP}(t)$ to eliminate the higher-frequency components, which gives the results of

$$x(t) = \frac{1}{2} \left(\text{Re}\{f(t)\} \cdot e^{j\zeta/2} \right. \quad (9)$$

$$+ \text{Im}\{f(t)\} \cdot e^{-j\zeta/2} \Big) \cdot e^{j\omega_c t} \otimes h_{LP}(t)$$

$$+ \frac{1}{2} \left(\text{Re}\{f(t)\} \cdot e^{j\zeta/2} \right.$$

$$- \text{Im}\{f(t)\} \cdot e^{-j\zeta/2} \Big) \cdot e^{-j\omega_c t} \otimes h_{LP}(t)$$

Note that the low-pass filter at the receiver always has a strict cutoff to control the noise level. The gains of the passband and stopband can significantly affect the extracted RFFs, especially the spectrum-domain RFFs. The low-pass filter function $h_{LP}(t)$ can be modeled as

$$H_{LP}(f) = \begin{cases} A_p, & |f| < W \\ A_s, & |f| > W \end{cases} \quad (10)$$

where A_p and A_s are the gains in the passband and stopband, respectively.

Finally, the received baseband signals $x(t)$ are sampled by analog-to-digital converter (ADC) to obtain the digital signal sequences which are sent to the identification units to extract the fingerprints. The ADC output is expressed as

$$x[n] = x(nT_s) + \Delta_n \quad (11)$$

where Δ_n is random quantization noise; for M-bit ADC quantization and input dynamic range $[-V, V]$, the maximum quantization error is $\delta_\Delta = 2^{-M} V$. Most existing RFF-based PLI solutions utilize the high-end measurement equipment (e.g., oscilloscope and spectrum analyzer) as the receiver. The key difference between the real wireless receivers lies in the sampling rate. While the oscilloscope and spectrum analyzer can directly sample the passband (or the baseband) signal at sampling rate in the order of GHz, most practical wireless device can only sample the baseband signal at the MHz rate.

It should be noted that the above signal reception and processing procedures are also subject to the influence of the hardware imperfections at the receiver's antenna, front-end, mixer, and ADC. However, those hardware imperfections are known to the receiver (or the identifier). Hence, the deviation or errors introduced by those hardware imperfections at the receiver can be adjusted and corrected when RFF is extracted.

RFF Extraction, Identification, and Classification

At the identification unit of the receiver, the RFFs are extracted from the sampled digital signals (The exception is the RFFs residing in the passband turn-on/off transient, which is digital sampled and extracted right after the RX front-end. Our model can be easily adjusted to describe such case by deleting the signal processing of mixer.) ($x[n]$ in (11)). The extracted feature can be expressed as

$$\mathbf{S} = \text{Feature}(\{x[n], n \in \mathbf{N}\}) \quad (12)$$

where $\mathbf{N}$ is the set of all sampled digital signals in this round of identification and $\text{Feature}(\cdot)$ is the feature extraction function. Depending on the properties of the RFF and the selected extraction strategy, the feature extraction function $\text{Feature}(\cdot)$ usually consists of domain change (e.g., Fourier transform, wavelet transform, and Hilbert transform, among others) and dimensionality reduction (e.g., linear discriminant analysis (LDA), principal component analysis (PCA), or simply picking us a certain parameter such as modulation errors and phase errors).

By now, the RFF $\mathbf{S}$ is obtained. The next step is to match the RFF with the reference fingerprint, $\mathbf{S}_R$, to calculate the distance matching scores and make the classification or identification decisions. The performance of the classification/identification can be evaluated by the error probabilities, which can be theoretically calculated through hypothesis testing models.

Classification

For N-users classification scenario, the N-hypothesis testing techniques can be adopted to apply to N genuine users with M_N reference fingerprints in the database for each user (Levine and Khuon 1992). The hypothesis $\mathcal{H}_N$ is that the obtained signal is from the genuine user $\#N$. Then the classification probability can be expressed by $P(\mathcal{H}_i | \mathcal{H}_j)$, $i, j = 1, \dots, N, i \neq j$, which is the probability that the RFF from genuine user $\#j$ is classified as the genuine user $\#i$. Such probability can be derived based on the feature distance between the extracted testing feature vector $\mathbf{S}$ with the reference feature vector $\mathbf{S}_R$:

$$D_i(\mathbf{S}) = \text{Distance}(\mathbf{S}, \mathbf{S}_R) \quad (13)$$

The testing feature vector $\mathbf{S}$ is matched with all the reference fingerprints and assigned to the identity with the smallest distance score. Then the classification error probability can be expressed as

$$P(\mathcal{H}_i | \mathcal{H}_j) = P(D_i = \min(\{D\}) | \mathcal{H}_j); \quad (14)$$

where the formulation of the probability $P(D_i = \min(\{D\}) | \mathcal{H}_j)$ depends on the distribution of $D_i(\mathbf{S})$. Then the average classification error rate can be derived:

$$P_e = \sum_{i=1}^N \sum_{j=1, j \neq i}^N P(\mathcal{H}_j | \mathcal{H}_i) P(\mathcal{H}_i) \quad (15)$$

Identification

Unlike the N -users classification, in the identification procedure, the distance scores between each authorized class are not the key metric. The whole network can be considered as a two-user hypothesis model (Danev and Capkun 2009; Polak et al. 2011; Polak and Goeckel 2011). $\mathcal{H}_1$ is the hypothesis that the incoming testing fingerprint is from the genuine users, while $\mathcal{H}_0$ indicates that the testing fingerprint is from an unknown imposter. In another word, all the N genuine users that have reference fingerprints stored in database can be treated as a whole class. A decision threshold λ is set up based on the calculated feature distance. If the minimal distance score is larger than the threshold, the testing feature is identified as an imposter's fingerprint, i.e., $\mathcal{H}_0$; otherwise it is judged as from a genuine user $\mathcal{H}_1$, i.e.,

$$\min(\{D\}) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\geq}} \lambda. \quad (16)$$

To evaluate the accuracy of identification performance, several probability metrics can be applied (Danev et al. 2009; Danev and Capkun 2009), including false accept rate (FAR) $P(\mathcal{H}_1 | \mathcal{H}_0)$, false reject rate (FRR) $P(\mathcal{H}_0 | \mathcal{H}_1)$, genuine accept rate (GAR) $P(\mathcal{H}_1 | \mathcal{H}_1)$, and genuine reject rate (GRR) $P(\mathcal{H}_0 | \mathcal{H}_0)$. Based on the decision rule given in (16), the above probabilities can be calculated in the same way as the classification case according to (13) and (14). We don't further elaborate the similar formulas here. To generate a single probability metric in identification system, many work put the above error rates together in the receiver operating characteristic (ROC) chart where the FRRs (or GARs) are plotted as a function of different FAR

levels. The operating point in ROC, where FAR and FRR are equal, is referred to as the equal error rate (EER).

Summary

The radio-frequency fingerprinting (RFF)-based physical layer identification (PLI) is a promising user device identification technique based on the inherent features from signal waveforms. In this entry, we rigorously model and discuss the sources of RFFs and device identification mechanisms along every single step in the RFF-based PLI techniques. The theoretical model provides the systematical description on the whole RFF-based PLI procedure, which is applicable in most state-of-the-art PLI techniques that are based on digital wireless communications.

References

- Brik V, Banerjee S, Gruteser M, Oh S (2008) Wireless device identification with radiometric signatures. In: Proceedings of the 14th ACM international conference on mobile computing and networking, ACM, pp 116–127
- D'Apuzzo M, D'Arco M, Liccardo A, Vadursi M (2010) Modeling DAC output waveforms. *IEEE Trans Instrum Meas* 59(11):2854–2862
- Danev B, Capkun S (2009) Transient-based identification of wireless sensor nodes. In: Proceedings of the 2009 international conference on information processing in sensor networks, IEEE Computer Society, pp 25–36
- Danev B, Heydt-Benjamin TS, Capkun S (2009) Physical-layer identification of RFID devices. In: Usenix security symposium, pp 199–214
- Danev B, Zanetti D, Capkun S (2012) On physical-layer identification of wireless devices. *ACM Comput Surv (CSUR)* 45(1):6
- Gharaibeh KM (2011) Nonlinear distortion in wireless systems: modeling and simulation with Matlab. Wiley
- Hall J, Barbeau M, Kranakis E (2004) Enhancing intrusion detection in wireless networks using radio frequency fingerprinting. In: Communications, internet, and information technology, pp 201–206
- Jana S, Kasera SK (2010) On fast and accurate detection of unauthorized wireless access points using clock skews. *IEEE Trans Mob Comput* 9(3):449–462

- Klein RW, Temple MA, Mendenhall MJ, Reising DR (2009) Sensitivity analysis of burst detection and RF fingerprinting classification performance. In: Communications, 2009. ICC'09. IEEE international conference on, IEEE, pp 1–5
- Levine RY, Khoun TS (1992) Training-set-based performance measures for data-adaptive decisioning systems. In: San Diego'92, International society for optics and photonics, pp 518–528
- Li Z, Xu W, Miller R, Trappe W (2006) Securing wireless systems via lower layer enforcements. In: Proceedings of the 5th ACM workshop on wireless security, ACM, pp 33–42
- Liu MW, Doherty JF (2008) Specific emitter identification using nonlinear device estimation. In: Sarnoff symposium, 2008 IEEE, IEEE, pp 1–5
- Polak AC, Goeckel DL (2011) RF fingerprinting of users who actively mask their identities with artificial distortion. In: Signals, systems and computers (ASILOMAR), 2011 conference record of the forty fifth asilomar conference on, IEEE, pp 270–274
- Polak AC, Dolatshahi S, Goeckel DL (2011) Identifying wireless users via transmitter imperfections. *IEEE J Sel Areas Commun* 29(7):1469–1479
- Rehman SU, Sowerby K, Coghill C (2012a) RF fingerprint extraction from the energy envelope of an instantaneous transient signal. In: Communications theory workshop (AusCTW), 2012 Australian, IEEE, pp 90–95
- Rehman SU, Sowerby K, Coghill C, Holmes W (2012b) The analysis of RF fingerprinting for low- end wireless receivers with application to IEEE 802.11 a. In: Mobile and wireless networking (iCOST), 2012 international conference on selected topics in, IEEE, pp 24–29
- Suski WC, Temple MA, Mendenhall MJ, Mills RF (2008) Using spectral fingerprints to improve wireless network security. In: Global telecommunications conference, 2008. IEEE GLOBE-COM 2008. IEEE, IEEE, pp 1–5
- Toonstra J, Kinsner W (1995) Transient analysis and genetic algorithms for classification. In: WES-CANEX 95. Communications, power, and computing. Conference proceedings., IEEE, IEEE, vol 2, pp 432–437
- Toonstra J, Kinsner W (1996) A radio transmitter fingerprinting system odo-1. In: Electrical and computer engineering, 1996. Canadian conference on, IEEE, vol 1, pp 60–63
- Xiao L, Greenstein L, Mandayam N, Trappe W (2007) Fingerprints in the ether: Using the physical layer for wireless authentication. In: Communications, 2007. ICC'07. IEEE international conference on, IEEE, pp 4646–4651
- Zanetti D, Danev B, et al (2010) Physical-layer identification of uhf RFID tags. In: Proceedings of the sixteenth annual international conference on mobile computing and networking, ACM, pp 353–364

Radiomap-Database

► Location Based Service of Ad Hoc Networks

Radio-on-Demand Wireless LAN

Suhua Tang¹ and Zhi Liu²

¹Department of Computer and Network Engineering, The University of Electro-Communications, Chofu, Japan

²Department of Mathematical and Systems Engineering, Shizuoka University, Hamamatsu, Japan

Synonyms

On-demand WLAN; Wake-on-wireless; Wake-on-WLAN; Wake-up control of WLAN

Definitions

Radio-on-demand wireless LAN (WLAN) defines a WLAN where each node (a power-hungry WLAN module) wakes up to communicate under a request, that is, each node is equipped with a low-power wake-up radio (WuR), which receives a wake-up message and activates its co-located WLAN module on demand.

Historical Background

Radio-on-demand WLAN realizes the wake-up control of a WLAN module. This function becomes urgent because nowadays WLAN modules are widely embedded into mobile devices (such as smartphones) driven by batteries with limited capacity.

A WLAN module can have different states with different power consumptions, as shown in

Table 1. The power relationship usually is Transmitting > Receiving > Idle > Transition > Sleep > Power-off. Without any power-save mechanism, a WLAN module is in the continuously awake mode (Idle, Receiving, or Transmitting). Analysis of traffic statistics indicates that WLAN traffic is bursty. A WLAN module is in the Idle state during most time, but its idle listening leads to much energy waste.

Power consumption of a WLAN module is highly related to the media access control (MAC) protocol design (Liu et al. 2018), and duty-cycling is a conventional method to this problem. In WLAN, duty-cycling is realized as the power-save mode. Each node periodically wakes up to receive traffic indication information from its associated access point (AP), by setting up a wake-up timer before entering the sleep state. But power consumption in the sleep state is non-negligible because some part like memory is still running, and periodical wake-up (transition from Sleep to an active state and vice versa) also causes much energy overhead. In the power-save mode, the longer a node sleeps, the more it saves power, but at the cost of a larger latency, which degrades quality of service (QoS).

To achieve both a long sleep and a fast response, an aggressive method is to completely turn off a WLAN module in its idleness to the Power-off state and activate it on demand. The remote wake-up control of a WLAN module from the Power-off state uses an out-of-band mechanism, which requires the cooperation between a sender and its target receiver. This was initially implemented and evaluated in Shih

et al. (2002), where an IEEE 802.15.4 radio is used as a WuR for the wake-up control of an IEEE 802.11b module. After a few years of research and development, currently, IEEE 802.11ba standards task group is working on an amendment of wake-up radio to enable low-power and low-latency WLANs (McCormick 2017). It should be noted that remote wake-up control alone is not always sufficient. A more aggressive policy is to harvest ambient energy to charge each node as well (Li et al. 2018).

Foundations

A WLAN module in the Idle state performs carrier sense for future channel access or receives packet headers to detect potential incoming packets. It consumes almost the same power as in the Receiving state. A WLAN module typically uses the same transceiver for both the control and data communications. In addition, power consumption of a WLAN transceiver increases with the supported data rate and the number of antennas. The control plane actually has little traffic, which does not require a high data rate. Therefore, it is necessary to separate the control plane from the data plane, at least in the wake-up control.

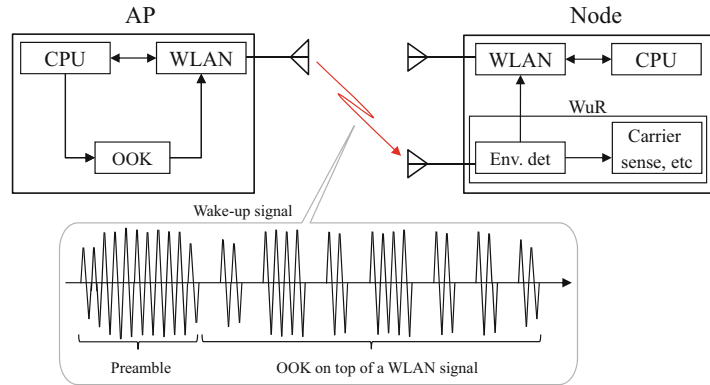
The key point of wake-up control is to equip each node with a low-power WuR for the wake-up signaling (at a low transmission rate), as shown in Fig. 1. A WuR can be implemented by different methods. (i) A generic low-power device, such as ZigBee or Bluetooth low energy (BLE), etc., can be used as a WuR (Shih et al.

Radio-on-Demand Wireless LAN, Table 1 States of a WLAN module

State	Meaning of the state
Power-off	A WLAN module is completely turned off, and its power consumption is nearly zero
Sleep	Although most part of a WLAN module is turned off, its memory (holding the associating state, etc.) is still working, whose power consumption is non-negligible
Transition	A WLAN module transits from the Sleep or Power-off state to an active state or vice versa
Idle	A WLAN module is active, but is neither transmitting a packet to others nor receiving a packet destined to itself. But it may overhear packets destined to other nodes
Receiving	A WLAN module is receiving a packet destined to itself
Transmitting	A WLAN module is transmitting a packet to other nodes

Radio-on-Demand Wireless LAN, Fig. 1

Integration of a wake-up radio with a WLAN module



2002). (ii) A specific device can be designed to only realize the wake-up function, whose power consumption can be reduced to be very low (Demirkol et al. 2009). A wake-up message is usually transmitted by the simple on-off-keying (OOK) and is received by envelope detection instead of coherent detection. In these methods, wake-up messages and WLAN signals are transmitted over different channels to avoid mutual interference.

A new trend is to transmit a wake-up signal on the same channel as a WLAN signal, for two reasons: (1) A WuR shares part of the RF circuit with a WLAN module to reduce the hardware cost. In addition, a node itself can use its WLAN module to transmit a wake-up message to activate other nodes, without requiring an extra transmitter. (2) A WuR is used for carrier sense and link quality detection, etc., besides the wake-up control. The wake-up control can be implemented by two methods: (i) Frame length of WLAN signals is adjusted to transmit a wake-up message from a WLAN module to a WuR (Chebrolu and Dhekne 2009; Tang et al. 2015). (ii) WLAN-emulated OOK is used to transmit a wake-up message by switching on/off inside a WLAN signal. In either case, a WuR performs envelope detection to receive a wake-up message, and the latter is more spectrum efficient.

Co-existence of Wake-Up Signals and WLAN Signals

When transmitting a wake-up signal on the same channel for WLAN signals, there is a problem of co-existence. In WLANs, continuous space

longer than DIFS (DCF inter-frame space) is usually regarded as an idle channel. To avoid a long idle period in an OOK-modulated signal, some mechanism, like Manchester coding, is applied (Tang et al. 2012). In addition, to further avoid interference, in the WLAN-emulated OOK, OOK is applied to the payload of a WLAN signal, but its header is kept unchanged, telling adjacent nodes when the wake-up signal ends (Park et al. 2016).

Unicast Wake-Up Versus Group Wake-Up

Wake-up control is usually used for the downlink traffic, where an AP activates a node when a packet is ready for the node. Using a unique wake-up ID helps to avoid false wake-up in this case. There is also broadcast traffic in a WLAN, but activating each node successively leads to much overhead. Therefore, a group wake-up is also necessary, where all nodes belonging to a group wake up together to receive a message from the AP.

Sometimes a selective wake-up among a group of nodes is also necessary. An AP without knowing the channel state information may try to activate a node whose link happens to be in the deep fading. A solution is to reuse a WuR for monitoring the channel. The AP sends a group wake-up ID. WuR of each node in the group monitors this wake-up message and, if its channel quality (SNR, signal-to-noise ratio) is above a threshold, contends to initiate its own reception (Tang and Obana 2017). This helps to improve the transmission rate and spectrum efficiency.

Wake-Up Latency

Due to the hardware constraint, a WLAN module, when receiving a wake-up request, cannot wake up immediately and has to transit from Power-off to the fully functional, active state. This transition period is called wake-up period and its length is the wake-up latency (Tang and Obana 2018). In the wake-up period, most time is for the high-frequency clock to get stable. This wake-up latency will cause a problem because the channel remains idle when it is expected to be busy, and some other node may interrupt with a new transmission, and the node in the wake-up transition will wake up to find the channel is already busy, which leads to a false-wake-up. In the unicast wake-up, this can be solved by letting the sender transmit a long preamble to hold the channel or transmit a control frame to reserve the channel.

The idleness per node has different patterns. In the case of long-term idleness, turning on a WLAN module from Power-off to the active state only has an initial delay. A loose integration of a WuR and a WLAN module works relatively well. On the other hand, to aggressively turn off WLAN modules in the short-term idleness (e.g., when other nodes are occupying the channel), a tight integration of a WuR and a WLAN module is necessary to reduce the impact of wake-up latency, which also requires a tight cooperation between a WLAN module and its associated AP.

Sensitivity

To reduce power consumption, a WuR uses a simple circuit of envelope detection without large power amplification. This affects the sensitivity of a WuR and the wake-up range. Fortunately, in the indoor environments with dense deployment of APs, the effective coverage of each AP is usually several meters, and the sensitivity problem of a WuR is not very serious.

Key Applications

The main application of radio-on-demand WLAN is in reducing the power consumption of mobile devices and future IoT devices that use the IEEE 802.11 series protocols.

Cross-References

- [Adaptive Medium Access Control for Internet-of-Things-Enabled MANETs](#)
- [QoS-Aware MAC](#)
- [Quality of Service in IEEE 802.11 Networks](#)

References

- Chebroly K, Dhekne A (2009) Esense: communication through energy sensing. In: Proceedings ACM MobiCom'09, pp 85–96
- Demirkol I, Ersoy C, Onur E (2009) Wake-up receivers for wireless sensor networks: benefits and challenges 16(4):88–96
- Li H, Ota K, Dong M (2018) Energy cooperation in battery-free wireless communications with radio frequency energy harvesting. *ACM Trans Embed Comput Syst* 17(2):44:1–44:17. <https://doi.org/10.1145/3141249>
- Liu Y, Ota K, Zhang K, Ma M, Xiong N, Liu A, Long J (2018) QTSAC: an energy-efficient MAC protocol for delay minimization in wireless sensor networks. *IEEE Access* 6:8273–8291. <https://doi.org/10.1109/ACCESS.2018.2809501>
- McCormick DK (2017) IEEE technology report on wake-up radio: an application, market, and technology impact analysis of low-power/low-latency 802.11 wireless LAN interfaces. 80211ba Battery Life Improvement: IEEE Technology Report on Wake-Up Radio, New York, pp 1–56. <https://doi.org/10.1109/IEEESTD.2017.8055459>
- Park M, Azizi S, Stacey R, Tian B, Vegt RD, Jones V, Au E, Yang DX, Yu RJ, Yang Y (2016) Proposal for wake-up receiver (WUR) study group. Technical Report document: IEEE 802.11-16/0722r1, IEEE
- Shih E, Bahl P, Sinclair MJ (2002) Wake on wireless: an event driven energy saving strategy for battery operated devices. In: Proceedings of the 8th annual international conference on mobile computing and networking, MobiCom'02. ACM, New York, pp 160–171. <https://doi.org/10.1145/570645.570666>
- Tang S, Obana S (2017) Energy efficient downlink transmission in wireless LANs by using low power wake-up radio. *Wirel Commun Mob Comput* 2017(2405381). <https://doi.org/10.1155/2017/2405381>
- Tang S, Obana S (2018) Reducing false wake-up in contention-based wake-up control of wireless LANs. *Wirel Netw*. <https://doi.org/10.1007/s11276-018-1662-y>
- Tang S, Yomo H, Kondo Y, Obana S (2012) Wake-up receiver for radio-on-demand wireless LANs. *EURASIP J Wirel Commun Netw* 2012:42
- Tang S, Yomo H, Takeuchi Y (2015) Optimization of frame length modulation-based wake-up control for green WLANs. *IEEE Trans Veh Technol* 64(2):768–780. <https://doi.org/10.1109/TVT.2014.2325643>

Random Access Technology in Cellular Networks

Maxime Grau, Chuan Heng Foh, and
Atta ul Quddus
5G Innovation Centre, Institute for
Communication Systems (ICS), University of
Surrey, Guildford, UK

Acronyms

1G	First generation
2G	Second generation
3G	Third generation
ACB	Access class barring
ACK	Acknowledgment
BS	Base station
C-RNTI	Cell radio network temporary identifier
CSMA	Carrier-sense multiple access
EC-GSM-IoT	Extended coverage GSM IoT
eMTC	Enhanced MTC
eNB	Evolved Node B
GSM	Global System for Mobile communications
H2H	Human-to-human
IoT	Internet of things
LTE	Long-Term Evolution
M2M	Machine-to-machine
MTC	Machine-type communication
NB-IoT	Narrowband IoT
PRACH	Physical random access channel
PUSCH	Physical uplink shared channel
RA	Random access
RAR	Random access response
RTD	Round-trip delay
TA	Timing alignment
UE	User equipment
UL	Uplink

Introduction

In cellular networks, a great number of users rely on a single centralized base station (BS) to communicate. Yet, base stations do not always

know when users may request access to the network, and they need to find ways to serve these unexpected users. This problem is addressed by a whole field of research, namely, random access (RA). There are multiple scenarios where RA can be used, e.g., first connection to the network, uplink (UL) scheduling request, timing alignment (TA) for non-synchronized users, positioning, etc. The most important scenario however, for which there exists no other method in wireless networks, is when the user is not known to the BS. On first connection, the BS does not expect the specific new user and has no other choice but to allocate specific resources that any user can use to make its presence known.

An inevitable consequence is that when multiple users compete for the same resource at the same time, the transmissions collide and cannot be decoded by the BS. Indeed, collisions can never be fully avoided, and RA techniques aim at reducing their probability while achieving an efficient trade-off between collision probability and resource usage. With the rise of the Internet of things (IoT), which relies on massive machine-type communication (MTC), random access protocols have seen a renewed interest in the recent years (Laya et al. 2014; Ijaz et al. 2016; Ferdouse et al. 2017) to cope with unprecedentedly high user density.

In this entry, an overview of the fundamental principles of random access and contention resolution is carried out in section “[Fundamentals of Random Access](#)”; how the main cellular communication systems implement these methods is studied in section “[Random Access in Practical Systems](#).” Finally, an outlook on the future of random access is discussed in section “[Future Outlook of Random Access](#).”

Fundamentals of Random Access

In this section, the two main families of random access are introduced. Methods for mitigating network congestion are then presented.

Radio Access Protocols

There exist two main categories of radio access: contention-based and reservation-based, both using random access to a certain amount.

Contention-Based

The first wireless RA system, ALOHAnet, was developed in the early 1970s by Norman Abramson's team of the University of Hawaii, using a **pure Aloha** protocol. Users would send their message at a random time on an uplink channel, and the base station would later broadcast messages it received so the users could know if their message collided with someone else's. Collision may appear when two or more transmissions appear at the same time. These transmissions may partly or fully interfere each other when they are initiated within each other vulnerable time. The vulnerable time can last for an overall of two transmission slots. Roberts proposed in 1972 (and later published in 1975) an improved version, namely **slotted Aloha** (Robert 1975), where RA slots were defined to limit the time that a transmission can be initiated. This proposal reduces the vulnerable time to one slot.

When a collision occurs, messages are assumed to be lost. This leads to high delay and low energy and spectral efficiency, as devices have to make multiple RA attempts. Ultimately, this can lead to network congestion.

Reservation-Based

For high traffic and long messages, contention-based RA is not a viable solution. To resolve this issue, modern wireless communication networks adopted a reservation-based RA, where each user gets assigned to a specific resource. As shown in Fig. 1, reservation-based protocols enable full utilization of the channel.

This can be achieved by using an additional handshake step, e.g., a preamble handshake. The principle of this method is to allocate a small portion of the spectrum to a contention-based step where users send a preamble. Preambles are very short messages mainly intended to make the user's presence known to the BS. The BS will then broadcast a RA response (RAR) to set up a dedicated connection with the user. There can be

multiple orthogonal preambles, meaning that the BS will detect multiple users who have randomly chosen different preambles.

This is particularly efficient for large messages and high traffic, preambles are indeed much smaller than messages, and the network can over-provision their number to reduce collision probability at this step. Collisions only lead to RA retry, which is more efficient than sending a whole message again.

A detailed example of a reservation-based RA system is presented in section "[RA in 4G LTE/LTE-Advanced](#)" with the Long-Term Evolution (LTE) RA procedure.

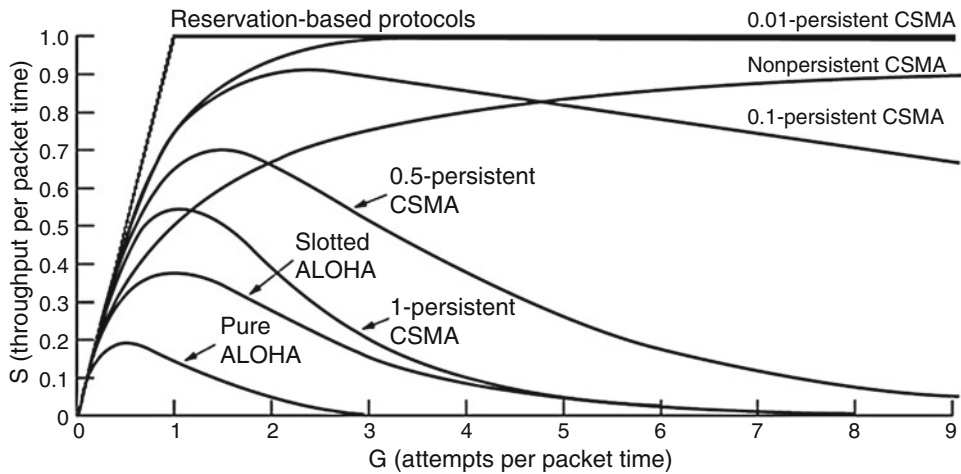
Collision Avoidance and Resolution

To address the collision issue and maximize throughput, techniques have been developed to take advantage of additional information.

Channel Sensing

In Aloha, users are unaware of their surrounding, and channel utilization cannot exceed $1/e$. To address this issue, Kleinrock and Tobagi (1975) proposed the **carrier-sense multiple access (CSMA)** protocol, which is contention-based and non-slotted. Its principle is that, before transmitting, a user senses if the channel is busy. In its purest form, namely, **1-persistent CSMA**, the user will not send its message as long as the channel is busy but will send it as soon as it detects that the channel is idle. In case two users were waiting for the current user to end, its transmission collides. An improvement of this technique is **p-persistent CSMA**, where users transmit with probability p as soon as they detect the channel is idle. Another variation is **nonpersistent CSMA**, where users stop sensing the channel if it is busy and wait a random amount of time before sensing it again. Channel throughput performance of these schemes is shown in Fig. 1.

This method can be adapted to slotted transmissions where messages are sent on multiple consecutive slots, during which the channel can be sensed as busy. This method performs best when the round-trip delay (RTD) between the users and the BS is negligible. Large cellular



Random Access Technology in Cellular Networks, Fig. 1 Random Access protocols Throughput (Kleinrock and Tobagi 1975)

networks, where RTD can reach orders a magnitude of 1 ms, cannot use this method. Therefore, protocols based on CSMA are only used in local networks such as Wi-Fi (IEEE Computer Society 2016).

Randomized Access Control

Reservation-based systems however use slotted Aloha to send preambles and cannot sense the channel. To prevent users from colliding over and over again as they all attempt RA on the next slot, users have to temporarily refrain from accessing the network. This can be achieved by using a **backoff** or a **barring** mechanism.

In the **random backoff** scheme, users wait for a random period of time after a collision before attempting RA again. In another variation, the **binary exponential backoff**, each time a user collides, it doubles the maximum random amount of time it can wait, thus doubling the expected amount of waiting time.

The **barring** scheme, on the other hand, adds another step before attempting RA again. Users have to randomly draw a number between 0 and 1 and compare it to the **persistence probability p**. If it is greater than p, they are allowed to attempt RA. Although both schemes look similar, barring distinguishes itself by allowing the barring parameter p to be modified and to be different for

different categories of users, as detailed in section “[RA for massive MTC](#).”

These two schemes are similar because they temporarily alleviate the load by preventing users from trying to access the network. Essentially, they both seek to optimally spread the traffic in time. These schemes are pushed to their limits with the **dynamic backoff** from Lin et al. (2016), while Rivero-Angeles et al. (2001) proposed the ancestor of **dynamic access class barring**. These schemes are essentially search algorithms that estimate the number of users to determine the optimized backoff window or barring parameter.

Collision-Free

To enable collision-free access, the users must be scheduled before they attempt to access the network. This can be achieved by using paging, where the BS calls out specific users. If they are listening, they may initiate a collision-free connection. For this scheme to be efficient however, the BS needs to make accurate guesses as to who is listening and needs resources.

Assuming perfectly random traffic, the BS cannot make any accurate guess, so users have to be very few and/or constantly scanning for a resource grant; otherwise a lot of downlink control resource is wasted. As a result, this method is either suited for very small networks (it is used in Bluetooth (Bluetooth Special Interest Group

2016)), delay-tolerant applications (users have to wait to be called to transmit uplink messages), and/or users with little to no energy constraints.

However, persistent scheduling (see section “[New techniques for Random Access](#)”) aims at using learning techniques to predict incoming traffic and allow collision-free access to a certain extent.

Random Access in Practical Systems

The history of RA follows that of multiple access: each generation of networks brings a radically different technology, and some of them can wait decades before being used. In this section, implementations of such technologies in standards are presented. Latest techniques for massive MTC are discussed.

RA in 1G Systems

Before cellular systems, the main radio access method was **push-to-talk**, so the RA was a pure Aloha protocol on a given channel. The first-generation (1G) system, Advanced Mobile Phone System (Tanenbaum and Wetherall 2011), brought access and control channels. Outgoing calls were done by (digitally) sending one’s and the receiver’s identifiers on a **pure Aloha** uplink control channel. The rest of the time, the devices were constantly listening to the downlink paging channel.

RA in 2G Systems

In the second-generation Global System for Mobile communications (GSM), a **slotted Aloha** reservation-based system was introduced. The GSM RA channel (ETSI 1996) is used to transmit a small message that contains the user’s identifier and can be used to request an uplink grant for a short message service, for instance. There is only one slot available, so if two or more users attempt RA at the same time, the BS will not be able to decode the resulting collided transmission and will not provide a RAR. Users then use **random backoff**. The BS also defines a maximum number of RA attempts.

RA in 3G Systems

The third generation of cellular networks Universal Mobile Telecommunications System (3GPP 2017a) uses a slotted Aloha reservation-based RA system. Preambles are coded with a sequence of 256 repetitions of Hadamard codes of length 16, thus making **16 orthogonal preambles**.

The user randomly chooses a signature from the set of available signatures within its Access Service Class, similar to **ACB** (see section “[RA for massive MTC](#)”), and sends it on the physical RA channel (PRACH). The BS responds on the Acquisition Indicator Channel by repeating the preamble and adding an acknowledgment (ACK).

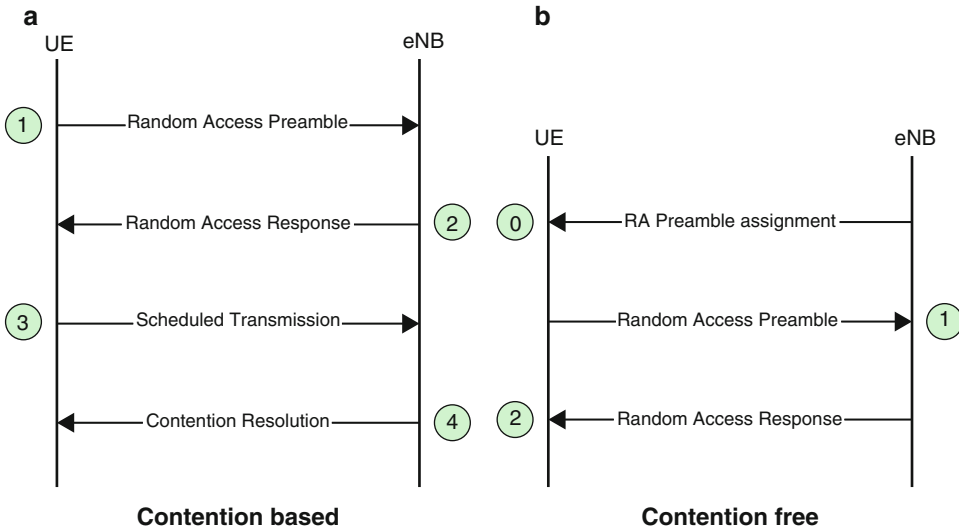
If the ACK is negative, the user performs **random backoff** with a predetermined time interval. If no ACK is received at the user equipment’s (UE) side, since the signature is not detected at base station, the user attempts RA again and ramps up in transmit power. Finally, if the ACK is positive, the user sends a RACH message (that includes user-specific connection setup information) in a dedicated slot matching the preamble signature. If two or more users sent the same preamble, the RACH messages will collide and will not be acknowledged. Collided users then proceed to random backoff.

RA in 4G LTE/LTE-Advanced

With the growing number of connected devices and the stringent synchronization requirements of the new multiplexing’s scheme, namely, orthogonal frequency-division multiplexing, RA fulfilled both the role of regular network access and also that of uplink synchronization, thanks to its preamble sequence’s properties.

LTE’s RA is a reservation-based system that uses **64 preambles**, 10 of which are reserved for contention-free RA (3GPP 2011), in a dedicated slotted Aloha control channel. These preambles are generated using a Zadoff-Chu sequence (Wen et al. 2006). Chu (1963, 1972) introduced these complex-polyphase sequences that have perfect autocorrelation properties, making them orthogonal by operating a cyclic shift.

LTE’s RA is a four-step procedure (see Fig. 2) with the following steps (3GPP 2016):



Random Access Technology in Cellular Networks, Fig. 2 RA procedure in LTE (3GPP 2017b)

- Step 1: The user equipment (UE) chooses a random preamble from the 54 orthogonal preambles and transmits it on the physical random access channel (PRACH).
- Step 2: The BS, named evolved node B (eNB), sends a random access response on the physical downlink shared channel that includes a TA command, a temporary cell radio network temporary identifier (C-RNTI), and an uplink grant on the physical uplink shared channel (PUSCH).
- Step 3: The UE transmits its identity using the C-RNTI to initiate a connection on the PUSCH.
- Step 4: If there was no collision in step 1, the eNB transmits a contention resolution message acknowledging the UE. Otherwise, UEs that collided attempt RA again.

Collisions happen at step 1. Users that collide will go through steps 2 and 3, and contention resolution cannot take place before step 4, resulting in additional delay and energy consumption for the UEs that collide. Note that steps 3 and 4 are not used in contention-free RA, e.g., for cell handover when the UE is already known by the network. Instead, it consists of three steps as the eNB assigns a preamble before step 1.

In case of a collision, users proceed to **binary exponential backoff**.

RA for Massive MTC

To prevent machine-to-machine (M2M) traffic from adversely affecting human-to-human (H2H) communication, **access class barring** (ACB) (3GPP 2012) is a standardized approach in LTE/LTE-Advanced. It assigns different persistence levels (the p probability, see section “Randomized Access Control”) to different traffic classes in order to ensure that high-priority users do not suffer from congestion caused by low-priority ones. **Dynamic RACH allocation** (3GPP 2011) is another technique proposed to alleviate the high RACH load problem in LTE, which is based on allocating additional RACH resources according to current traffic load conditions. Finally, **extended access barring** is the extreme case where certain users are completely barred from accessing the network ($p = 0$) as long as the BS judges that RACH congestion is too high.

LTE Release 13 introduced new features specifically aimed at enabling massive MTC (GSMA 2016): extended coverage GSM IoT (EC-GSM-IoT), enhanced MTC (eMTC),

and narrowband IoT (NB-IoT). The main focus however is to offer increased range and battery. eMTC improves on LTE with additional features such as extended idle-mode discontinuous reception and semi-persistent scheduling. These techniques enable the BS to learn traffic MTC devices' traffic patterns, thus improving **load prediction**. NB-IoT is essentially a simplified narrower-band LTE and comes with a new PRACH, the narrowband PRACH; this essentially means that RA is dedicated to MTC devices, contrarily to ACB that restricts their access to the PRACH.

Future Outlook of Random Access

Although it is critical to ever-growing multiple access networks, RA in standards has known little to no major evolution since slotted Aloha was introduced in the 1970s. Efforts were made to prevent congestion and maintain slotted Aloha's performances, but that is as far as current systems go. However, with the IoT becoming a reality, this should change rapidly. In this section, recent technologies are presented, and the future of RA is discussed.

New Requirements for 5G and the IoT

The massive MTC scenario represents a paradigm switch from traditional wireless networks. It relies on an unprecedented number (Osseiran et al. 2014) of low-energy consumption devices that are expected to be inactive most of the time (Qi et al. 2015), may be far from the BS, and sporadically transmit small messages with bursts of traffic peaks (3GPP 2011).

While, in LTE, UL requests are usually handled by the physical uplink control channel (PUCCH) for already connected UEs, they can only be made through RA for MTC devices waking up. This, along with the number of users, M2M traffic will undoubtedly overload the RACH. And while LTE was designed with enough overprovisioning to achieve a 99% RA success probability, it is commonly accepted that

a higher collision probability is expected for the IoT scenario (Lo et al. 2011).

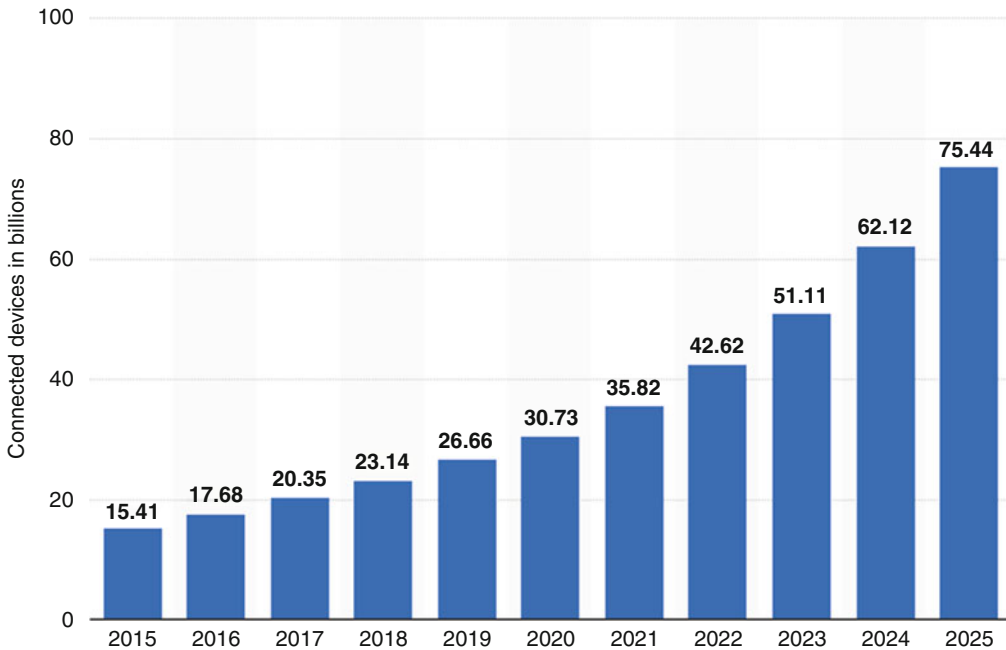
Although spectacular predictions in the early 2010s of a trillion devices by 2020 have since deflated (Nordrum 2016), around 30 billion devices are expected to be connected by 2020 and 75 billion by 2025 (see Fig. 3). It must be noted that the H2H market has already reached maturity in most part of the world and that most of the additional devices are expected to be MTC devices, while they are a minority today.

Another challenge to future RA systems is the burstiness of the traffic. Event-driven communication may lead to high traffic bursts that can pose a big threat to network access. A traffic burst can be modeled by a beta distribution (3GPP 2011) where the traffic peak is temporarily much higher than the average expected traffic. This has been shown to lead to congestion in traditional LTE networks.

New Techniques for Random Access

Many groundbreaking RA techniques have been proposed in the recent years and are good candidates to be part of future cellular systems. Some of the most promising methods are:

- **Coded RA:** Some new techniques focus on improving slotted Aloha's spectral efficiency. Wunder et al. (2015) proposed to use sets of preambles as a code, effectively increasing the overall spectral efficiency and theoretically achieving 100% spectrum usage.
- **Virtual preambles RA:** Techniques to virtually increase the number of preambles have been proposed. For instance, Munir et al. (2014) propose to send multiple RAR (with different UL grants) to any received preamble. If users collide at step 1 of the LTE procedure, they still have a chance to be served at step 3.
- **Non-orthogonal RA:** A revolutionizing approach is to design a pure contention-based RA, as opposed to reservation-based, where collisions can be resolved. Shirvanimoghaddam et al. (2017) propose such a scheme, which uses successive interference



Random Access Technology in Cellular Networks, Fig. 3 Number of connected devices in billions by 2025 (Nordrum 2016)

cancellation to resolve collisions and shows promising results for moderate traffic loads.

- Semi-persistent scheduling: MTC devices are expected to have a sporadic traffic load and be in sleep mode most of the time. Such traffic can be memorized by the BS. Qi et al. (2015) propose a semi-persistent paging mechanism by predicting when the MTC devices will be awake. This can reduce the RA load by allowing some collision-free access.
- Learning-based RA: Although it is assumed that traffic is purely random, users in a given cell will not always change. This is especially true for MTC devices that will mostly be stationary (e.g., sensors). Thus, it is possible for users to learn to share RA resources more efficiently. Bello et al. (2016) used Q-learning to achieve 100% RACH throughput by having users to converge to dedicated RACH slots.

Conclusion

RA in current cellular networks relies on well-established technologies that have proven to be

sufficient for H2H communication. However, the IoT will change the game in the upcoming years: there will be more devices, these devices will be more heavily dependent on RA, and they will access the network in bursts of traffic.

Random access, by definition, works under the assumption that the network is not able to predict when devices will need to access the network. However this does not need to always be the case. Nowadays, machine learning, traffic prediction, and context-aware mechanisms are gaining more and more traction. Indeed, one can argue that a certain amount of the upcoming massive MTC traffic, if not a significant amount, can be predicted and served by using alternative methods.

The next stepping stone of random access will be to only be used for communications that are truly random.

References

- 3GPP (2011) Study on RAN Improvements for Machine-type Communications
- 3GPP (2012) TS 36.331 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification

- 3GPP (2016) TS 36.300 V13.4.0 Radio Access Network (E-UTRAN); Overall description
- 3GPP (2017a) ETSI TS 125 211 V14.0.0 Universal Mobile Telecommunications System (UMTS); Physical channels and mapping of transport channels onto physical channels (FDD). 0
- 3GPP (2017b) TR 138 912 – V14.0.0 Study on New Radio (NR) access technology
- Bello LM, Mitchell P, Grace D, Mickus T (2016) Q-learning based random access with collision free RACH interactions for cellular M2M. In: Proceedings of NGMAST 2015 9th international conference next generation mobile applications, services and technologies, pp 78–83. <https://doi.org/10.1109/NGMAST.2015.22>
- Bluetooth Special Interest Group (2016) Bluetooth Core Specification Version 5.0. Bluetooth Core Specif Version 50 2684
- Chu DC (1963) Polyphase codes with good nonperiodic correlation properties. *IEEE Trans Inf Theory* 9:43–45. <https://doi.org/10.1109/TIT.1963.1057798>
- Chu D (1972) Polyphase codes with good periodic correlation properties (corresp.). *IEEE Trans Inf Theory* 18:531–532
- ETSI (1996) Digital cellular telecommunications system (Phase 2+); Mobile radio interface layer 3 specification (GSM 04.08)
- Ferdouse L, Anpalagan A, Misra S (2017) Congestion and overload control techniques in massive M2M systems: a survey. *Trans Emerg Telecommun Technol* 28:e2936. <https://doi.org/10.1002/ett.2936>
- GSMA (2016) 3GPP low power wide area technologies (LPWA). GSMA White Paper, 19
- IEEE Computer Society (2016) [802.11u] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications
- Ijaz A, Zhang L, Grau M et al (2016) Enabling massive IoT in 5G and beyond systems: PHY radio frame design considerations. *IEEE Access* 4:3322–3339. <https://doi.org/10.1109/ACCESS.2016.2584178>
- Kleinrock L, Tobagi F (1975) Packet switching in radio channels: part I – carrier sense multiple-access modes and their throughput-delay characteristics. *IEEE Trans Commun* 23:1400–1416. <https://doi.org/10.1109/TCOM.1975.1092768>
- Laya A, Alonso L, Alonso-Zarate J (2014) Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives. *IEEE Commun Surv Tutor* 16:4–16. <https://doi.org/10.1109/SURV.2013.111313.00244>
- Lin G-Y, Chang S-R, Wei H-Y (2016) Estimation and adaptation for Bursty LTE random access. *IEEE Trans Veh Technol* 65:2560–2577. <https://doi.org/10.1109/TVT.2015.2418811>
- Lo A, Member S, Law YW, Jacobsson M (2011) Enhanced LTE-advanced random-access mechanism for massive machine-to-machine (M2M) communications. In: 27th meet wireless world research forum, pp 1–7
- Munir D, Hasan SF, Chung MY, Kim JS (2014) Enhancement of LTE RACH through extended random access process. *Electron Lett* 50:1399–1400. <https://doi.org/10.1049/el.2014.1159>
- Nordrum A (2016) Popular internet of things forecast of 50 billion devices by 2020 is outdated. *IEEE Spectr*. <https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated>
- Osseiran A, Boccardi F, Braun V et al (2014) Scenarios for 5G mobile and wireless communications: the vision of the METIS project. *IEEE Commun Mag* 52:26–35. <https://doi.org/10.1109/MCOM.2014.6815890>
- Qi Y, Quddus AU, Ali Imran M, Tafazolli R (2015) Semi-persistent RRC protocol for machine-type communication devices in LTE networks. *IEEE Access* 3:864–874. <https://doi.org/10.1109/ACCESS.2015.2445976>
- Rivero-Angeles ME, Lara-Rodriguez D, Cruz-Perez FA (2001) A new EDGE medium access control mechanism using adaptive traffic load slotted ALOHA. In: *IEEE 54th vehicular technology conference. VTC fall 2001. Proceedings (Cat. No.01CH37211)*, IEEE, pp 1358–1362
- Roberts LG (1975) ALOHA packet system with and without slots and capture. *SIGCOMM Comput. Commun.* 5(2):28–42. <https://doi.org/10.1145/1024916.1024920>
- Shirvanimoghaddam M, Condoluci M, Dohler M, Johnson SJ (2017) On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT. *IEEE J Sel Areas Commun* 35:2238–2252. <https://doi.org/10.1109/JSAC.2017.2724442>
- Tanenbaum AS, Wetherall D (2011) *Computer networks*, 5th edn. Pearson Prentice Hall, Boston
- Wen YWY, Huang WHW, Zhang ZZZ (2006) CAZAC sequence and its application in LTE random access. In: *2006 IEEE information theory Workshop – ITW '06 Chengdu*, pp 544–547. <https://doi.org/10.1109/ITW2.2006.323692>
- Wunder G, Stefanovic C, Popovski P, Thiele L (2015) Compressive coded random access for massive MTC traffic in 5G systems. In: *2015 49th Asilomar conference on signals, systems and computers*. IEEE, pp 13–17

Random Distance Distribution

Fei Tong¹ and Jianping Pan²

¹The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

²University of Victoria, Victoria, BC, Canada

Synonyms

Geometrical probability; Distance distribution; Wireless networks

Problem Definition

As a significant complementary tool to the Poisson point process and binomial point process models in stochastic geometry, probabilistic distance-based models have been extensively studied and applied for interference analysis in finite wireless networks (Tong and Pan 2017; Tong et al. 2017a). Such models eventually involve distributions of the distances between transmitting and receiving nodes, called nodal distance distributions (NDDs) in this entry. NDDs intrinsically depend on network coverage and nodal spatial distribution. Depending on whether one of the nodes is given or not, there are two types of NDDs. One is Ref2Ran NDD from a given reference node to a random node and the other is Ran2Ran NDD between two random nodes. Both of them have been widely utilized for interference analysis in wireless networks, as shown in Ahmadi et al. (2013, 2015), Tong et al. (2017b), Tong and Pan (2017), and the references therein. How to effectively obtain the relevant NDDs is therefore significant to interference modeling and analysis in wireless networks. We have surveyed the state-of-the-art approaches in Tong et al. (2017a), as also briefly shown and extended below.

Key Results

Ref2Ran NDD

A systematic recursive approach has been proposed in Ahmadi and Pan (2014) to obtain the Ref2Ran NDD from an arbitrary reference node to a random node inside an arbitrary polygon. Firstly, the Ref2Ran NDD from a vertex of an arbitrary triangle to a random node inside the triangle is obtained by using the area-ratio approach, based on which the Ref2Ran NDD from an exterior or interior reference node to a random node inside the triangle is then obtained by using a decomposition and recursion (D&R) method. In this entry, an NDD from an arbitrary reference node to a random node inside an arbitrary triangle is the so-called Ref2Ran triangle-NDD.

Since any polygon can be triangulated, the Ref2Ran NDDs associated with an arbitrary polygon can be obtained based on the obtained Ref2Ran triangle-NDDs. Specifically, if the reference node R is inside the polygon, the polygon can be triangulated from R as shown in Fig. 1a; otherwise, if R is outside, the triangle can be triangulated from a vertex as shown in Fig. 1b. Let N denote the number of generated triangles after the triangulation ($N = 6$ and 4 for the polygons shown in Fig. 1a, b, respectively), S_i^Δ the area of triangle $\mathcal{T}_i$, S the area of the polygon, and $F_i^\Delta(d)$ the cumulative distribution function (CDF) of the distance from R to a random node inside the triangle $\mathcal{T}_i$ (which has already been obtained). Then, with a weighted probabilistic sum, the CDF of the distance from R to a random node inside the polygon is

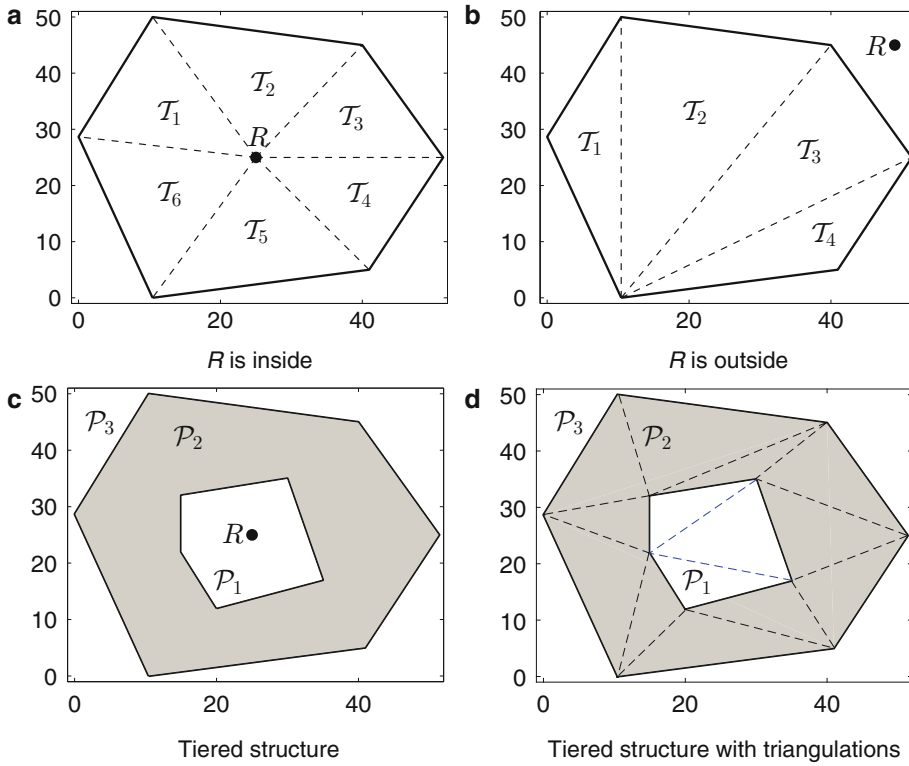
$$F(d) = \sum_{i=1}^N \frac{S_i^\Delta}{S} F_i^\Delta(d). \quad (1)$$

It is possible that different subareas of a network have different node densities (the so-called tiered network). For this case, the above weighted probabilistic sum approach in (1) is still applicable but with different weights due to the node density difference. Take the network shown in Fig. 1c, for example, and suppose the node density ratio between $\mathcal{P}_1$ (the white area) and $\mathcal{P}_2$ (the grey area) is $\lambda_1 : \lambda_2$ (λ_1 and λ_2 are not zero simultaneously). Let S_i denote the area of $\mathcal{P}_i$ and $F_i(d)$ the CDF of the distance from R to a random node inside $\mathcal{P}_i$ (which has been obtained in (1)). Then the CDF of the distance from an arbitrary reference node R to a random node inside the whole area $\mathcal{P}_3$ ($\mathcal{P}_1$ plus $\mathcal{P}_2$) is

$$F_3(d) = \sum_{i=1}^2 \frac{S_i \lambda_i}{\sum_j S_j \lambda_j} F_i(d). \quad (2)$$

Ran2Ran NDD

For Ran2Ran NDDs, we recently proposed an approach in Tong and Pan (2017), which can handle the networks in the shape of arbitrary polygons as well as the polygons with different



Random Distance Distribution, Fig. 1 An arbitrary reference node R and arbitrary polygons (Tong et al. 2017a). (a) R is inside. (b) R is outside. (c) Tiered structure. (d) Tiered structure with triangulations

node densities in different subareas. The D&R method is also utilized to this end through polygon triangulation. Specifically, Ran2Ran NDDs associated with arbitrary triangles (the so-called Ran2Ran triangle-NDDs) are firstly obtained, which include the Ran2Ran NDD within an arbitrary triangle and that between two arbitrary triangles. Then Ran2Ran NDDs associated with arbitrary polygons can be obtained through the D&R method. We refer interested readers to Tong and Pan (2017) for more details.

Nonuniform node distribution is also considered in the above approach. Take the irregular polygon with a triangulation as shown in Fig. 1b, for example. Assume the node density ratio among $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, and $\mathcal{T}_4$ is $\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4$ ($\lambda_1, \lambda_2, \lambda_3$, and λ_4 are not zero simultaneously). Through D&R, the CDF of the Ran2Ran NDD within the polygon is given by a weighted probabilistic sum,

$$F(d) = \sum_{i=1}^4 \sum_{j=1}^4 \frac{S_i^{\Delta} \lambda_i S_j^{\Delta} \lambda_j}{\left(\sum_{k=1}^4 S_k^{\Delta} \lambda_k \right)^2} F_{ij}^{\Delta}(d), \quad (3)$$

where S_x^{Δ} is the area of triangle $\mathcal{T}_x$, and $F_{ij}^{\Delta}(d)$ is the CDF of the Ran2Ran NDD within triangle $\mathcal{T}_i$ if $i = j$, or between two triangles $\mathcal{T}_i$ and $\mathcal{T}_j$ if $i \neq j$, which have been obtained above as the Ran2Ran triangle-NDDs.

For tiered polygons shown as in Fig. 1c, the CDFs of the Ran2Ran NDDs within $\mathcal{P}_1$ and $\mathcal{P}_2$, i.e., $F_{11}(d)$ and $F_{22}(d)$, respectively, can be obtained by (3). Assume the node density ratio among $\mathcal{P}_1$ and $\mathcal{P}_2$ is $\lambda_1 : \lambda_2$ (λ_1 and λ_2 are not zero simultaneously). To obtain the CDF of the Ran2Ran NDD between $\mathcal{P}_1$ and $\mathcal{P}_2$, i.e., $F_{12}(d)$, we first triangulate each of them as shown in Fig. 1d and get 3 and 11 triangles, respectively. Then with a weighted probabilistic sum,

$$F_{12}(d) = \sum_{m=1}^3 \sum_{n=1}^{11} \frac{S_m^{\Delta_1} S_n^{\Delta_2}}{\sum_{i=1}^3 S_i^{\Delta_1} \sum_{j=1}^{11} S_j^{\Delta_2}} F_{mn}^{\Delta}(d), \quad (4)$$

where $S_x^{\Delta_y}$ is the area of the x -th triangle in $\mathcal{P}_y$ and $F_{mn}^{\Delta}(d)$ is the CDF of the Ran2Ran NDD between the m -th triangle in $\mathcal{P}_1$ and the n -th triangle in $\mathcal{P}_2$. With $F_{11}(d)$, $F_{22}(d)$, and $F_{12}(d)$ obtained above, $F_{33}(d)$, $F_{13}(d)$, and $F_{23}(d)$ can be obtained by

$$F_{33}(d) = \sum_{i=1}^2 \sum_{j=1}^2 \frac{S_i \lambda_i S_j \lambda_j}{\left(\sum_{k=1}^2 S_k \lambda_k\right)^2} F_{ij}(d), \quad (5)$$

$$F_{3i}(d) = \sum_{j=1}^2 \frac{S_j \lambda_j}{\sum_{k=1}^2 S_k \lambda_k} F_{ij}(d), \quad (i \in \{1, 2\})$$

where S_x is the area of $\mathcal{P}_x$, and $F_{ij}(d)$ is the CDF of the Ran2Ran NDD within $\mathcal{P}_i$ if $i = j$, or between $\mathcal{P}_i$ and $\mathcal{P}_j$ if $i \neq j$.

Interference Analysis

Assume a general path loss model. The path loss ratio of the transmission power at distance d is

$$L(d) = \beta d_0^\alpha d^{-\alpha}, \quad (6)$$

where β is a path loss constant determined by the hardware features of transceivers, d_0 is a given reference distance, and α is the path loss exponent. As a result, the received signal strength at distance d is

$$P_r(d) = L(d)P_t, \quad (7)$$

where P_t is the transmission power. Therefore, given a distance distribution, the distributions of path loss ratio and received signal strength can also be obtained by using the change-of-variable technique. For all the transmitters in the network, assume that the i -th transmitter has a transmission power of P_i . The cumulative interference at the receiver from all its interfering nodes is

$$I = \sum_i P_i L(d_i), \quad (8)$$

where $L(d_i)$ is given in (6) and d_i is the distance from the receiver to the i -th interfering node. Correspondingly, the distribution of the interference at the receiver can also be obtained based on distance distributions.

Key Applications

The interference model introduced above can be utilized to analyze the distributions of some important performance metrics in wireless networks, such as signal-to-interference-plus-noise ratio (SINR), link capacity, etc. As shown in (8), the key is to obtain the distributions of the relevant distances. Some examples are introduced below.

In Zhuang et al. (2011), the NDDs from a BS (reference node) to an arbitrary cellular user (random node) in the same hexagonal cell or to the users in its adjacent cells were derived. Based on the obtained NDDs, the received exterior co-channel interference at a BS from a transmitting cellular user of any six adjacent cells was explicitly analyzed. In Ahmadi et al. (2013), the obtained NDDs were further utilized to analyze the cumulative exterior interference at a BS from its all first-layer neighbor cells. In Ahmadi et al. (2015), the distance distributions associated with arbitrarily shaped cells in tiered cellular systems were utilized to model and analyze a cellular network consisting of a single macrocell and multiple randomly deployed femtocells in a link reusing scenario. The distributions of interference were derived, based on which the system outage probability was obtained. In Tong et al. (2017b), we proposed a framework using the probabilistic distance model introduced above to obtain the distributions of the signal and interference strengths, and further SINR, based on which, the performance metrics that are functions of the SINR could be analyzed, such as outage probability and capacity.

References

- Ahmadi M, Pan J (2014) Random distances associated with arbitrary triangles: a recursive approach with an arbitrary reference point. In: UVicSpace, pp 1–13. <http://hdl.handle.net/1828/5134>
- Ahmadi M, Ni M, Pan J (2013) A geometrical probability-based approach towards the analysis of uplink inter-cell interference. In: Proceedings of the IEEE global communications conference (GLOBECOM), pp 4952–4957
- Ahmadi M, Tong F, Zheng L, Pan J (2015) Performance analysis for two-tier cellular systems based on probabilistic distance models. In: Proceedings of the IEEE international conference on computer communications (INFOCOM), pp 352–360
- Tong F, Pan J (2017) Random-to-random nodal distance distributions in finite wireless networks. *IEEE Trans Veh Tech* 66(11):10070–10083
- Tong F, Pan J, Zhang R (2017a) Distance distributions in finite ad hoc networks: approaches, applications, and directions. In: *Ad hoc networks*, pp 167–179
- Tong F, Wan Y, Zheng L, Pan J, Cai L (2017b) A probabilistic distance-based modeling and analysis for cellular networks with underlaying device-to-device communications. *IEEE Trans Wirel Commun* 16(1):451–463
- Zhuang Y, Luo Y, Cai L, Pan J (2011) A geometric probability model for capacity analysis and interference estimation in wireless mobile cellular systems. In: Proceedings of the IEEE global communications conference (GLOBECOM), pp 1–6

Random Motion

- [Brownian Motion](#)

Random Walk

- [Brownian Motion](#)

Reachability Management

- [Mobility Management in 5G](#)

Reaction-Diffusion

- [Reaction-Diffusion Channels](#)

Reaction-Diffusion Channels

Yansha Deng
Department of Informatics, King's College
London, London, UK

Synonyms

[Chemical reaction](#); [Enzyme reaction](#); [First-order reaction](#); [Mathematical model](#); [Particle-based simulation](#); [Reaction-diffusion](#); [Second-order reaction](#)

Definitions

In a practical biological system, chemical reactions widely occur during the molecule propagation environments; thus it is very important to understand the channel response of the molecular communication system under various reaction-diffusion channels. Historically, chemical reactions describe the changes only involved in the positions of electrons in the forming and breaking of chemical bonds between atoms, with no change to the nuclei. There are basically four types of reactions categorized in terms of their functionalities, namely, synthesis reactions, decomposition reactions, oxidation and reduction reactions, and complexation reactions. This chapter focuses on the mathematical models and simulation approaches for the reaction-diffusion channels under the first-order reaction, second-order reaction, and enzyme reaction.

Mathematical Models for Reaction-Diffusion Channel

The simplest propagation method in molecular communication is free diffusion, where the

molecules' random motion is governed by the Brownian motion (Berg 1993). In a practical biological system (e.g., human body), numerous complex chemical reactions take place in the human body. The general reaction-diffusion equation for species S is Debnath (2005, Eq. (8.12.1))

$$\frac{\partial C_S}{\partial t} = D_S \nabla^2 C_S + f(C_S, \vec{r}, t), \quad (1)$$

where $f(\cdot)$ is the reaction term.

First-Order Reaction

With the emission of A information molecules at the transmitter, the fundamental mechanism of the first-order reaction can be described as

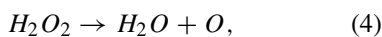


and k_F is the forward reaction rate constant with units in $\text{molecule}^{-1}\text{m}^3\text{s}^{-1}$.

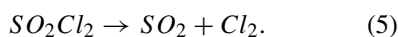
The general reaction-diffusion equation for the first-order reaction is written as

$$\begin{aligned} \frac{\partial C_A}{\partial t} &= D_A \nabla^2 C_A - k_F C_A, \\ \text{and } \frac{\partial C_B}{\partial t} &= D_B \nabla^2 C_B + k_F C_A, \end{aligned} \quad (3)$$

where D_A and D_B are the constant diffusion coefficients of the A and B molecules and C_A and C_B are the time-varying concentrations of the A and B molecules, which can be solved using (3) with the defined initial and boundary conditions. The channel responses of molecular communication system with the first-order reaction have been studied in Chou (2015), Heren et al. (2015), Ahmadzadeh et al. (2016), and Deng et al. (2017a). The first-order reaction examples include

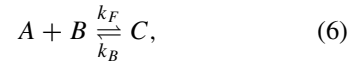


and



Second-Order Reaction

With the emission of A information molecules at the transmitter, the fundamental mechanism of the second-order reaction can be described as

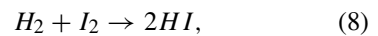


and k_B is the backward reaction rate constant with units in s^{-1} .

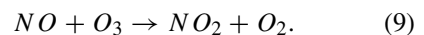
The general reaction-diffusion equation for the second-order reaction is written as

$$\begin{aligned} \frac{\partial C_A}{\partial t} &= D_A \nabla^2 C_A - k_F C_A C_B + k_B C_C, \\ \frac{\partial C_B}{\partial t} &= D_B \nabla^2 C_B - k_F C_A C_B + k_B C_C, \\ \text{and } \frac{\partial C_C}{\partial t} &= D_C \nabla^2 C_C - k_B C_C + k_F C_A C_B, \end{aligned} \quad (7)$$

where D_C is the constant diffusion coefficient of the C molecules and C_A , C_B , and C_C are the time-varying concentrations of the A , B , and C molecules, which can be solved using (7) with the defined initial and boundary conditions. The tractable analytical solutions for the second-order reaction are usually not easy to obtain, and the channel responses of molecular communication system with the second-order reaction have been studied in Chou (2015), Ahmadzadeh et al. (2016), and Deng et al. (2017b). The first-order reaction examples include



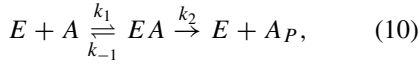
and



Enzyme Reaction

In a practical biological system, the biological catalysts (i.e., enzymes) take part in various chemical reactions. Motivated by this, the enzymatic reactions in the propagation environments were studied in Noel et al. (2014) for intersymbol interference (ISI) reduction. With the emission of A information molecules at the transmitter, the

fundamental mechanism of enzymatic reactions can be described via Michaelis-Menten kinetics as



where E is the enzyme molecule, EA is the intermediate, A_P is the degraded A molecule, and k_1 , k_{-1} , and k_2 are the reaction rate constants with units in molecule⁻¹m³s⁻¹, s⁻¹, and s⁻¹.

To evaluate the channel response of the molecular communication system, we describe the spatiotemporal behavior of the information, enzyme, and intermediate molecules via reaction-diffusion partial differential equations as

$$\begin{aligned} \frac{\partial C_A}{\partial t} &= D_A \nabla^2 C_A - k_1 C_A C_E + k_{-1} C_{EA}, \\ \frac{\partial C_E}{\partial t} &= D_E \nabla^2 C_E - k_1 C_A C_E + k_{-1} C_{EA} \\ &\quad + k_2 C_{EA}, \\ \frac{\partial C_{EA}}{\partial t} &= D_A \nabla^2 C_{EA} + k_1 C_A C_E - k_{-1} C_{EA} \\ &\quad - k_2 C_{EA}. \end{aligned} \quad (11)$$

Solving (11) along with the initial and boundary conditions defined at the emission and the reception of the molecular communication system, the time-varying concentration of each type of molecules can be obtained. Note that solving (11) in a closed form is analytically intractable (or at least very difficult) for the boundary conditions that are commonly considered in MC (point source, preexisting enzyme, etc.).

Simulation Approaches for Reaction-Diffusion Channel

The particle-based simulations can be generally categorized as the mesoscopic and microscopic models (Noel et al. 2017). In mesoscopic model, the spaces are divided into many subvolumes (i.e., cubes), where the number of molecules in each subvolume is individually counted (Chou 2015). In microscopic model, molecules are tracked individually in each simulation step

(Deng et al. 2016; Ahmadzadeh et al. 2016). Here, we limit ourselves to the microscopic simulation model due to its simplicity, efficiency, and generality.

Diffusion

The time is divided into small simulation intervals of size Δt , and each time instant is $t_m = m\Delta t$, where m is the current simulation index. According to Brownian motion, the displacement of a molecule in each dimension in one simulation step Δt can be modeled by an independent Gaussian distribution with variance $2D\Delta t$ and zero mean $\mathcal{N}(0, 2D\Delta t)$. The displacement ΔS of a molecule in a 3D fluid environment in one simulation step Δt is therefore

$$\Delta S = \{\mathcal{N}(0, 2D\Delta t), \mathcal{N}(0, 2D\Delta t), \mathcal{N}(0, 2D\Delta t)\}. \quad (12)$$

In each simulation step, the number of molecules and their locations are stored. Note that the accuracy of the simulation results depends on the size of simulation step Δt .

First-Order Reaction

The first-order reaction in the microscopic regime can usually be simulated using the probability (Andrews 2009, Eq. (22))

$$P_f = 1 - e^{-k_F \Delta t}, \quad (13)$$

where Δt is the simulation step.

Second-Order Reaction

With two reactants A and B , we need to carefully determine the binding radius r_{AB} . When A and B are physically separated and the binding radius r_{AB} is much larger than the expected separation of the A and B molecules due to diffusion in the next simulation time step, the binding will occur in the next time step. Note that the root-mean-square of the separation of the A and B molecules in a single simulation time step is Andrews and Bray (2004, Eq. (23))

$$r_{\text{rms}} = \sqrt{2(D_A + D_B)\Delta t}. \quad (14)$$

If Δt is sufficiently large, then $r_{\text{rms}} \gg r_{AB}$, and the binding radius r_{AB} can be defined as Andrews and Bray (2004, Eq. (27))

$$r_{AB} = \left(\frac{3k_F \Delta t}{4\pi} \right)^{\frac{1}{3}}. \quad (15)$$

When $r_{\text{rms}} \gg r_B$ is not satisfied, the binding radius r_{AB} can be evaluated using numerical methods (Andrews and Bray 2004).

The backward reaction probability can be characterized using

$$P_b = 1 - e^{-k_B \Delta t}. \quad (16)$$

Enzyme Reaction

The enzyme reaction includes the second-order reactions; thus the binding radius of the binding reaction k_1 can be described as

$$r_{EA} = \left(\frac{3k_1 \Delta t}{4\pi} \right)^{\frac{1}{3}}, \quad (17)$$

when $r_{\text{rms}} \gg r_{EA}$, otherwise r_{EA} can be obtained using numerical methods (Andrews and Bray 2004).

The probabilities of the unbinding reactions occurring are a function of both the unbinding and degradation rate constants, written as Andrews and Bray (2004, Eq. (14))

$$\Pr(\text{Reaction } k_{-1}) = \frac{k_{-1}}{k_{-1} + k_2} [1 - \exp(-\Delta t (k_{-1} + k_2))], \quad (18)$$

and

$$\Pr(\text{Reaction } k_2) = \frac{k_2}{k_{-1} + k_2} [1 - \exp(-\Delta t (k_{-1} + k_2))], \quad (19)$$

respectively.

Cross-References

- [Brownian Motion](#)
- [Modulation in Molecular Signaling](#)
- [Modeling approaches for simulating molecular communications](#)
- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

References

- Ahmadzadeh A, Arjmandi H, Burkovski A, Nallanathan A, Schober R (2016) Comprehensive reactive receiver modeling for diffusive molecular communication systems: reversible binding, molecule degradation, and finite number of receptors. *IEEE Trans NanoBiosci* 15:713–727
- Andrews SS (2009) Accurate particle-based simulation of adsorption, desorption, and partial transmission. *Phys Biol* 6:046015
- Andrews SS, Bray D (2004) Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Phys Biol* 1:137
- Berg HC (1993) Random walks in biology. Princeton University Press, Princeton
- Chou CT (2015) Impact of receiver reaction mechanisms on the performance of molecular communication networks. *IEEE Trans Nanotech* 14:304–317
- Debnath L (2005) Nonlinear partial differential equations for scientists and engineers. Birkhaeuser, Boston
- Deng Y, Noel A, Elakashlan M, Nallanathan A, Cheung KC (2016) Modeling and simulation of molecular communication systems with a reversible adsorption receiver. *IEEE Trans Mol Biol Multi-Scale Commun* 1:347–362
- Deng Y, Guo W, Noel A, Nallanathan A, Elakashlan M (2017a) Enabling energy efficient molecular communication via molecule energy transfer. *IEEE Commun Lett* 21:254–257
- Deng Y, Pierobon M, Nallanathan A (2017b) A microfluidic feed forward loop pulse generator for molecular communication, in *Proc. of IEEE GLOBECOM'17*, Singapore, pp 1–7
- Heren AC, Yilmaz HB, Chae CB, Tugcu T (2015) Effect of degradation in molecular communication: impairment or enhancement? *IEEE Trans Mol Biol Multi-Scale Commun* 1:217–229
- Noel A, Cheung KC, Schober R (2014) Improving receiver performance of diffusive molecular communication with enzymes. *IEEE Trans NanoBiosci* 13:31–43
- Noel A, Cheung KC, Schober R, Makrakis D, Hafid A (2017) Simulating with AcCoRD: actor-based communication via reaction–diffusion. *Nano Commun Netw* 11:1878–7789

Received Signal Model

- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

Received Spatial Correlation

- [Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System](#)

Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion

Arman Ahmadzadeh, Vahid Jamali, Wayan Wicke, and Robert Schober
Institute for Digital Communications,
Friedrich-Alexander University
Erlangen-Nürnberg (FAU), Erlangen, Germany

Synonyms

[Channel impulse response](#); [Received signal model](#); [Reception mechanisms](#)

Definitions

Receiver mechanisms refer to the set of processes which describe how signaling molecules in the molecular communication environment are sensed by a nanoreceiver.

Historical Background

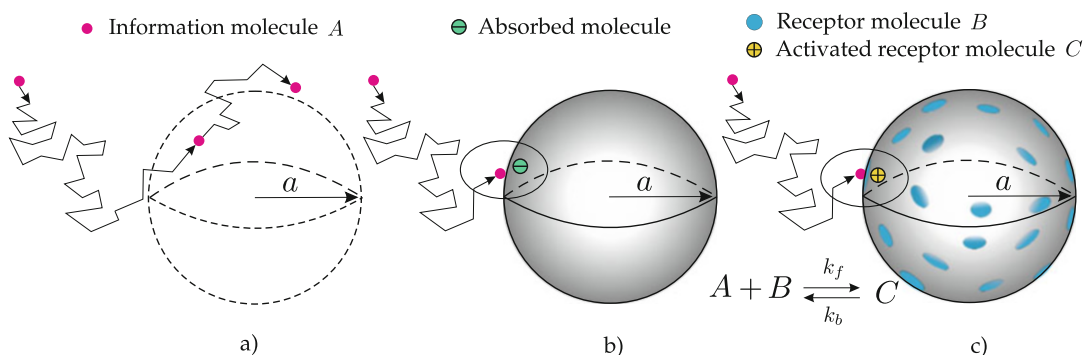
Molecular communication (MC) is a potentially bio-compatible approach for enabling commu-

nication at micro- to nanoscale. In fact, MC is one of the foremost means of communication between biological entities in nature. The bio-compatibility and abundant use of MC in nature make it an attractive candidate for synthetic MC system design.

Meaningful yet simple models that take the physics of the components of synthetic MC systems (i.e., the transmitter, channel, and receiver) at nanoscale into account are a crucial first step towards synthetic MC system design. In particular, in this entry, a review of the most common receiver models in the MC literature is presented.

In nature, typically, a biological entity such as a cell measures the concentration of signaling molecules in its vicinity via the receptor protein molecules embedded in its surface. These measurements are inherently *stochastic* and corrupted by various noise sources, namely, (1) the random movements of the signaling molecules diffusing from the source to the cell, (2) stochastic degradation and/or production of signaling molecules during diffusion, (3) stochastic binding and unbinding of signaling molecules to the receptor protein molecules on the surface of the cell, and (4) the stochastic nature of the signaling pathways relaying the noisy measurements of the receptors from the exterior to the interior of the cell. Based on the noisy estimation of the concentration of the signaling molecules, the cell decides on a suitable response, see ten Wolde et al. (2016). In the following, noise sources (1), (2), (3), and (4) are referred to as random walk (RW), degradation and production (D&P), receptor kinetic (RK), and signaling pathway (SP) noise sources, respectively.

In the MC literature, different receiver models have been proposed, each taking some of the noise sources mentioned above into account. In particular, in many works, such as Pierobon and Akyildiz (2011a), Mahfuz et al. (2014), a receiver model with only RW noise, the so-called *passive receiver*, is considered. A receiver model considering RW and D&P noise is proposed in Noel et al. (2014). In Pierobon and Akyildiz



Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion, Fig. 1 Schematic presentation of three common receiver models: a) passive receiver, b) fully absorbing receiver, and c) reactive receiver

(2011b), ShahMohammadian et al. (2013), RW and RK noise are considered, while the diffusion process and the reaction mechanisms describing RK noise are treated separately. The so-called *fully absorbing receiver* model is proposed in Yilmaz et al. (2014). There, the impact of RW noise and a special case of RK noise are jointly considered. In particular, for RK noise, it is assumed that (I) the surface of the receiver is uniformly covered with infinitely many receptors and (II) the reactions corresponding to RK noise are modeled deterministically, i.e., it is assumed that any collision of a signaling molecule and the receiver surface leads to absorption of the signaling molecule. Extensions of the *fully absorbing receiver* model taking D&P noise into account and relaxing assumption (I) are reported in Heren et al. (2015) and Akkaya et al. (2015), respectively. A receiver model which includes RW and SP noise for two simple approximate signaling pathway models is proposed in (Chou 2015). A receiver model that jointly takes the impact of RW and RK noise with assumption (I) into account is proposed in Deng et al. (2015). A comprehensive model for *reactive receivers* is introduced in Ahmadzadeh et al. (2016), where the joint effects of RW, D&P, and RK noise are considered. This entry reviews the receiver signal models for *passive*, *fully absorbing*, and *reactive* receivers in *diffusive* MC systems, see Fig. 1. Most of the other above-mentioned receiver models are special cases or simple extensions of these three models.

Key Applications

These receiver models can be applied for the design and optimization of synthetic MC systems, from modulation schemes for the transmitter to detection and/or decoding algorithms for the receiver.

Receiver Signal Model

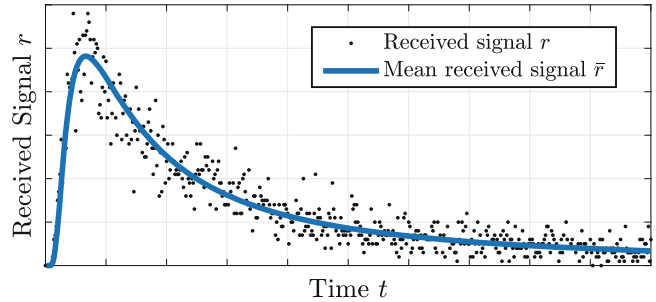
As discussed in the previous section, the received signal is highly random and corrupted by several types of noise. The *random* received signal is denoted by r , its distribution by $f(r)$, and its mean and variance by $\bar{r}$ and σ_r^2 , respectively. The quantitative meaning of r depends on the particular choice of the receiver and will be discussed in detail in the following subsections. Furthermore, this entry focuses on spherical receivers with radius a [m] and the model of r for *one impulsive* release of N identical information (or signaling) A molecules into an unbounded fluid environment at time $t_0 = 0$ by a point transmitter located at a distance d [m] from the center of the receiver. The diffusion coefficient of the A molecules is denoted by D [m²/s]. A summary of the relevant system parameters is provided in Table 1. An example for the typical shape of the mean received signal $\bar{r}$ in diffusive MC systems is depicted in Fig. 2. In particular, when $N = 1$, $\bar{r}$ is referred to as the *channel impulse response* of the considered MC system and denoted by $\bar{h}$, i.e.,

Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion, Table 1 Summary of system parameters

Symbol	Definition	Symbol	Definition
a	Receiver radius	D	Diffusion coefficient of A molecules
V	Volume of the receiver hull	k_f	Binding reaction rate constant
d	Distance between transmitter and receiver	k_b	Unbinding reaction rate constant
N	# of impulse release of A molecules	k_d	Degradation reaction rate constant
M	# of receptors on the surface of receiver	r_s	Radius of circular receptor patch

Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion, Fig. 2

Example for sample realization and mean of a typical received signal in diffusive MC systems



$\bar{h} = \bar{r}_{|N=1}$. In other words, $\bar{h}$ is in fact the *probability* of receiving one given A molecule. In the following, first, a review of the existing models for $\bar{h}$ is given for the passive, fully absorbing, and reactive receivers. Subsequently, a discussion is provided on the common assumptions adopted in the literature to arrive at mathematically tractable expressions for $\bar{r}$ in the case of $N > 1$. At the end, this entry review some of the existing models proposed for the distribution of r , $f(r)$.

Channel Impulse Response

Passive Receiver

In the passive receiver model, the randomness of r is caused only by the diffusion of the signaling molecules, where the receiver does not influence the propagation of the A molecules. In particular, signaling A molecules in the vicinity of the receiver can enter and leave the receiver via free diffusion, see e.g., Fig. 1a). The passive receiver model is a good approximation for the diffusion of small, uncharged molecules, such as ethanol, urea, and oxygen. These molecules can enter and leave a cell by passive diffusion across the plasma membrane, (Alberts et al. 2009). In the passive receiver model, r is defined as the number of

observed A molecules inside the receiver hull. It has been shown in Noel et al. (2013), Eq. (27) that $\bar{h}$ is given by

$$\bar{h} = \frac{1}{2} \left(\operatorname{erf} \left(\frac{a-d}{\sqrt{4Dt}} \right) + \operatorname{erf} \left(\frac{a+d}{\sqrt{4Dt}} \right) \right) + \frac{\sqrt{Dt}}{a\sqrt{\pi}} \left(\exp \left(-\frac{(a-d)^2}{4Dt} \right) + \exp \left(-\frac{(a+d)^2}{4Dt} \right) \right), \quad (1)$$

where $\operatorname{erf}(\cdot)$ denotes the error function. A common approach, which is referred to as the *uniform concentration assumption*, is to approximate $\bar{h}$ everywhere inside the volume of the receiver, $V = 4/3\pi a^3$, by its value at the center of the receiver. This leads to the following simple expression for $\bar{h}$

$$\bar{h} = \frac{y}{(4\pi Dt)^{3/2}} \exp \left(-\frac{d^2}{4Dt} \right). \quad (2)$$

It was shown in Noel et al. (2013) that (2) provides an accurate approximation for (1) if $a < 0.15 d$ holds.

Fully Absorbing Receiver

For the fully absorbing receiver, the signaling A molecules that reach the receiver via diffusion are absorbed as soon as they hit the receiver surface,

see Fig. 1b). In this case, r is defined as the number of absorbed molecules during infinitesimally small time dt . It is shown in Yilmaz et al. (2014) that the number of absorbed molecule until time t , g , can be obtained as

$$g = \frac{a}{d} \operatorname{erfc}\left(\frac{d-a}{\sqrt{4Dt}}\right), \quad (3)$$

where $\operatorname{erfc}(\cdot)$ denotes the complementary error function. Given (3), $\bar{h}$ can be written as (Yilmaz et al. 2014, Eq. (22))

$$\bar{h} = dg = \frac{a(d-a)}{td\sqrt{4\pi Dt}} \exp\left(-\frac{(d-a)^2}{4Dt}\right) dt. \quad (4)$$

Remark 1 Alternatively, when the receiver counts the number of absorbed molecules during a window of $t_u - t_l$ seconds, $\bar{h}$ can be defined as

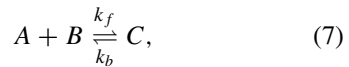
$$\begin{aligned} \bar{h} &= g|_{t_u} - g|_{t_l} \\ &= \int_{t_l}^{t_u} \frac{a(d-a)}{\tau d\sqrt{4\pi D\tau}} \exp\left(-\frac{(d-a)^2}{4D\tau}\right) d\tau. \end{aligned} \quad (5)$$

Reactive Receiver

Reactive receiver is a good model for large or polar molecules that cannot passively diffuse through the hull of a cell and have to be detected by external receptors embedded in the surface of the cell. For the reactive receiver model, RW, D&P, and RK noise sources jointly affect r . In particular, the diffusive signaling A molecules can potentially degrade in the channel and those which reach the receiver participate in a reversible bimolecular second-order reaction with receptor protein B molecules on the surface of the receiver to activate receptors and form the ligand-receptor complex C molecules, see e.g., Fig. 1c). The first-order degradation reaction in the channel and the reversible reaction on the receiver surface are given by



and



respectively. Here, k_d , k_f , and k_b denote the degradation reaction rate constant in $[s^{-1}]$, the binding reaction rate constant in $[\text{molecule}^{-1} \cdot \text{m}^3 \cdot \text{s}^{-1}]$, and the unbinding reaction rate constant in $[s^{-1}]$, respectively. For the reactive receiver, r is defined as the number of activated receptors C . The expression describing $\bar{h}$ is derived in (Ahmadzadeh et al. 2016, Eq. (29))

$$\begin{aligned} \bar{h} &= \frac{k_f^* e^{-k_d t}}{4\pi da\sqrt{D}} \left\{ \frac{\alpha W\left(\frac{d-a}{\sqrt{4Dt}}, \alpha\sqrt{t}\right)}{(\gamma - \alpha)(\alpha - \beta)} \right. \\ &\quad + \frac{\beta W\left(\frac{d-a}{\sqrt{4Dt}}, \beta\sqrt{t}\right)}{(\beta - \gamma)(\alpha - \beta)} \\ &\quad \left. + \frac{\gamma W\left(\frac{d-a}{\sqrt{4Dt}}, \gamma\sqrt{t}\right)}{(\beta - \gamma)(\gamma - \alpha)} \right\}, \end{aligned} \quad (8)$$

where $W(n, m) = \exp(2nm + m^2)\operatorname{erfc}(n + m)$, and constants α , β , and γ are the solutions to the following system of equations:

$$\begin{cases} \alpha + \beta + \gamma = \left(1 + \frac{k_f^*}{4\pi aD}\right) \frac{\sqrt{D}}{a}, \\ \alpha\gamma + \beta\gamma + \alpha\beta = k_b - k_d, \\ \alpha\beta\gamma = k_b \frac{\sqrt{D}}{a} - k_d \left(1 + \frac{k_f^*}{4\pi aD}\right) \frac{\sqrt{D}}{a}. \end{cases} \quad (9)$$

Here, k_f^* is defined as

$$k_f^* = \frac{4\pi D a k_f \varphi}{k_f (1 - \varphi) + 4\pi D a}, \quad (10)$$

where φ is a constant that is given by (Ahmadzadeh et al. 2016, Eq. (39))

$$\varphi = \frac{Mr_s^2 (k_f + 4\pi Da)}{a(1 - \lambda)(\pi r_s k_f + 16\pi Da^2) + Mr_s^2 (k_f + 4\pi Da)}, \quad (11)$$

and $\lambda = M \frac{\pi r_s^2}{4\pi a^2}$ is the fraction of the receiver surface covered by M circular receptors, each with radius r_s .

Remark 2 The fully absorbing receiver is a special case of the reactive receiver when $\lambda \rightarrow 1$, $k_b = k_d = 0$, and $k_f \rightarrow \infty$. In this case, reaction equation (7) becomes a pseudo first-order reaction of the form $A \rightarrow C$, with binding reaction rate constant $k_f \rightarrow \infty$, where now C denotes the number of absorbed molecules. However, $k_f \rightarrow \infty$ implies that any collision of a signaling A molecule with the receiver surface leads to the formation of a C molecule, i.e., the reaction is deterministic. It can be shown that under these assumptions (8) simplifies to (3).

Remark 3 The expressions for $\bar{h}$ of the passive and fully absorbing receivers can be extended to take *first-order* degradation reactions in the channel of the form in (6) into account by multiplying the corresponding $\bar{h}$ with $\exp(-k_d t)$. For higher-order enzymatic degradation reactions in the channel, we refer the reader to Noel et al. (2014).

Stochastic Received Signal

In order to calculate the mean and the distribution of r when $N \geq 1$, i.e., $\bar{r}$ and $f(r)$, respectively, the following set of assumptions are commonly made (explicitly or implicitly) in the MC literature. For evaluation of $\bar{r}$, if one assumes that (a) *each* of the RW, D&P, RK, and SP noise sources affects individual A molecules *independently*, i.e., if one assumes that the diffusion, reaction, and reception processes of a given A molecule do not influence the diffusion, reaction, and reception processes of any other A molecule, $\bar{r}$ is obtained as

$$\bar{r} = N\bar{h}. \quad (12)$$

For calculation of $f(r)$, in addition to assumption (a), if one further assumes that (b) the transmitter releases *exactly* N A molecules, and (c) the reception mechanism at the receiver can be described as a binary process (i.e., a molecule

is either counted or it is not counted), then, r follows a Binomial distribution with mean $\bar{r}$ and variance $\sigma_r^2 = \bar{r}(1 - \bar{h})$. The latter assumption is valid for all three considered receiver models. For example, for the passive receiver (fully absorbing receiver) at any time t , a given A molecule is either inside the volume of the receiver (absorbed) or not. Similarly, for the reactive receiver, at any time t , a receptor B molecule is either activated (i.e., a C molecule is formed) or it is not activated. Thus, $f(r)$ can be written as follows

$$\begin{aligned} r &\sim \text{Binomial}(N, \bar{h}), f(r) \\ &= \binom{N}{r} (\bar{h})^r (1 - \bar{h})^{N-r}. \end{aligned} \quad (13)$$

The Binomial distribution complicates the analysis of MC systems. As a result, two approximations of the Binomial distribution are commonly used for analysis:

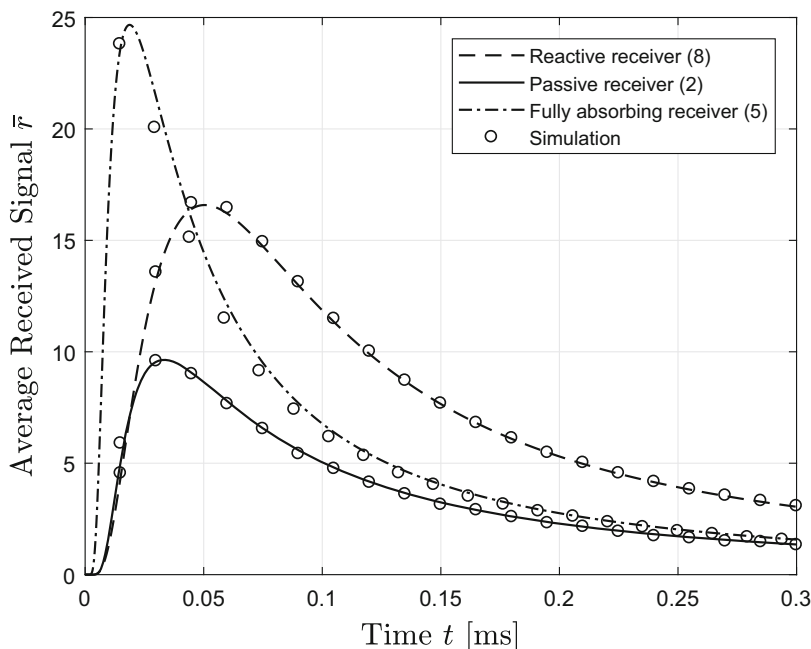
- **Gaussian Distribution:** Under the assumption that $\bar{r}$ is sufficiently large for the central limit theorem to apply, r can be modeled as a Gaussian random variable with mean $\bar{r}$ and variance $\sigma_r^2 = \bar{r}(1 - \bar{h})$, i.e.,

$$\begin{aligned} r &\sim \mathcal{N}(\bar{r}, \bar{r}(1 - \bar{h})), \\ f(r) &= \frac{1}{(2\pi\bar{r}(1 - \bar{h}))^{1/2}} \exp\left(-\frac{(r - \bar{r})^2}{2\bar{r}(1 - \bar{h})}\right). \end{aligned} \quad (14)$$

The Gaussian distribution is much more amenable to analysis than the Binomial distribution. It is worth mentioning that the accuracy of the Gaussian distribution relies on basic assumption that $\bar{r}$ is large, i.e., the received signal is strong.

- **Poisson Distribution:** The Binomial distribution can be approximated by a Poisson distribution if N is very large and $\bar{h}$ is very small. Both of these assumptions usually hold for diffusive MC systems. Thus, we can write

$$r \sim \text{Poiss}(\bar{r}), \quad f(r) = \frac{\bar{r}^r \exp(-\bar{r})}{r!}. \quad (15)$$



Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion, Fig. 3 $\bar{r}$ versus time t for three different receiver models

Results and Discussions

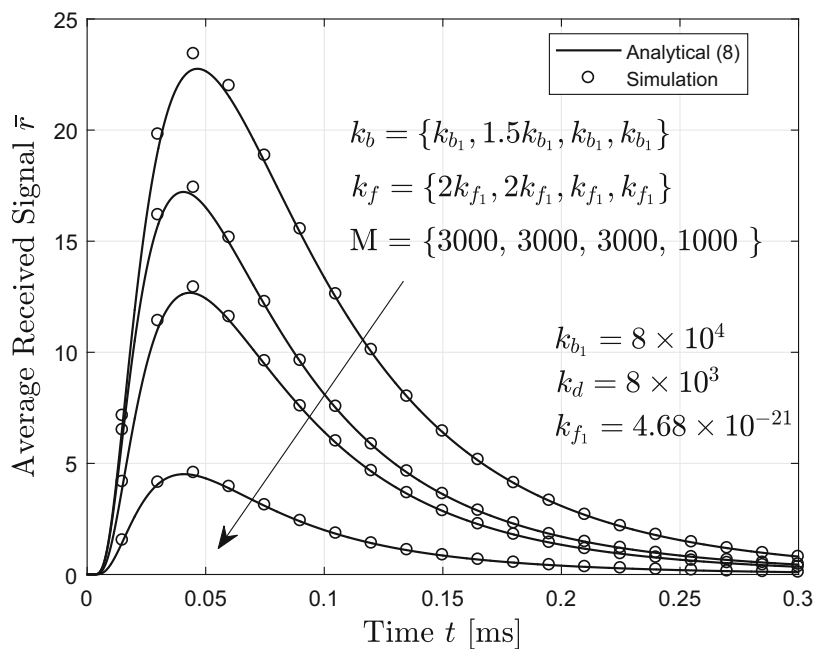
For all simulation results, the following set of system parameters is considered: $d = 1$ [μm], $a = 0.25$ [μm], $N = 2000$, and $D = 5 \times 10^{-9}$ [m^2/s]. Moreover, the simulation results are obtained via standard particle-based simulation of Brownian motion.

Figure 3 shows $\bar{r}$ versus time t for the three considered receiver models. For the fully absorbing receiver, (5) with $t_u = t + \Delta t$, $t_l = t$, and $\Delta t = 6$ [μs] is employed. Furthermore, for the reactive receiver, $k_f = 4.68 \times 10^{-21}$ [$\text{molecule}^{-1} \cdot \text{m}^3 \cdot \text{s}^{-1}$] and $k_b = 8 \times 10^4$ [s^{-1}] are chosen. For all curves $k_d = 0$. Simulation results confirm the accuracy of the expressions reviewed in the previous sections. It can also be observed that all curves have similar shapes, i.e., first $\bar{r}$ increases, then it decreases.

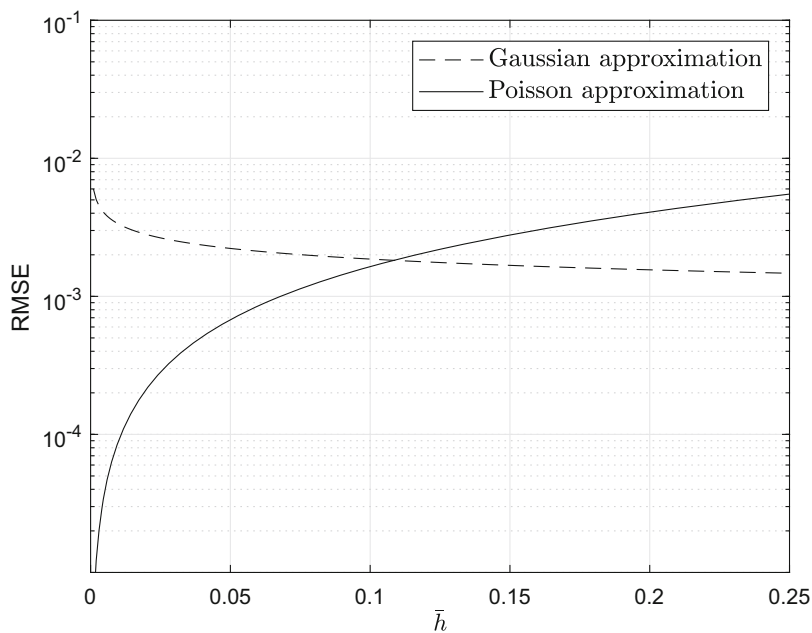
Figure 4 shows $\bar{r}$ for the reactive receiver versus time t for different system parameters; $M = \{3000, 1000\}$, $k_f = \{4.68, 9.36\} \times 10^{-21}$ [$\text{molecule}^{-1} \cdot \text{m}^3 \cdot \text{s}^{-1}$], $k_b = \{8, 12\} \times 10^4$ [s^{-1}], and $k_d = 8 \times 10^3$ [s^{-1}]. Figure 4 shows that,

for given k_f and k_b , employing more receptors M increases $\bar{r}$. This was expected as increasing M increases the chance that a receptor is activated. It can also be observed that, for given k_b and M , a larger value of k_f increases $\bar{r}$ as the capture probability of the signaling molecules is increased. Furthermore, for given k_f and M , an increase in the value of k_b decreases $\bar{r}$, since for larger values of k_b , the unbinding reaction is more likely to happen and, as a result, more signaling molecules can dissociate and escape from the vicinity of the receiver.

In Fig. 5, the root mean squared error (RMSE) between the Gaussian/Poisson cumulative distribution function (CDF) and the Binomial CDF is plotted vs. $\bar{h}$. Figure 5 reveals that for increasing $\bar{h}$, the accuracy of the Poisson model deteriorates, whereas the accuracy of the Gaussian model improves. This behavior was expected from the assumptions needed for the validity of the approximate models. However, for typical diffusive MC systems, $\bar{h} \ll 0.1$ holds. As a result, the Poisson model is generally more accurate.



Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion, Fig. 4 $\bar{r}$ for reactive receiver vs. time t for different receiver parameters



Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion, Fig. 5 RMSE versus $\bar{h}$ between the approximated Gaussian/Poisson CDF and the Binomial CDF

References

- Ahmadzadeh A, Arjmandi H, Burkovski A, Schober R (2016) Comprehensive reactive receiver modeling for diffusive molecular communication systems: reversible binding, molecule degradation, and finite number of receptors. *IEEE Trans Nanobioscience* 15(7): 713–727
- Akkaya A, Yilmaz HB, Chae CB, Tugcu T (2015) Effect of receptor density and size on signal reception in molecular communication via diffusion with an absorbing receiver. *IEEE Commun Lett* 19(2):155–158
- Alberts B, Bray D, Hopkin K, Johnson AD, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2009) *Essential Cell Biology*, 3rd edn. Garland Science, New York
- Chou CT (2015) Impact of receiver reaction mechanisms on the performance of molecular communication networks. *IEEE Trans Nanotechnol* 14(2):304–317
- Deng Y, Noel A, Elakashan M, Nallanathan A, Cheung KC (2015) Modeling and simulation of molecular communication systems with a reversible adsorption receiver. *IEEE Trans Mol, Biol Multi-Scale Commun* 1(4):347–362
- Heren AC, Yilmaz HB, Chae CB, Tugcu T (2015) Effect of degradation in molecular communication: impairment or enhancement? *IEEE Trans Mol, Biol Multi-Scale Commun* 1(2):217–229
- Mahfuz M, Makrakis D, Mouftah H (2014) A comprehensive study of sampling-based optimum signal detection in concentration-encoded molecular communication. *IEEE Trans Nanobioscience* 13(3):208–222
- Noel A, Cheung KC, Schober R (2013) Using dimensional analysis to assess scalability and accuracy in molecular communication. In: *Proceedings of the IEEE international conference on communications*, pp 818–823
- Noel A, Cheung KC, Schober R (2014) Improving receiver performance of diffusive molecular communication with enzymes. *IEEE Trans Nanobioscience* 13(1):31–43
- Pierobon M, Akyildiz I (2011a) Diffusion-based noise analysis for molecular communication in nanonetworks. *IEEE Trans Signal Process* 59(6):2532–2547
- Pierobon M, Akyildiz I (2011b) Noise analysis in ligand-binding reception for molecular communication in nanonetworks. *IEEE Trans Signal Process* 59(9):4168–4182
- ShahMohammadian H, Messier G, Magierowski S (2013) Modelling the reception process in diffusion-based molecular communication channels. In: *Proceedings of the IEEE international conference on communications*, pp 782–786
- ten Wolde PR, Becker NB, Ouldrige TE, Mugler A (2016) Fundamental limits to cellular sensing. *J Stat Phys* 162(5):1395–1424
- Yilmaz H, Heren A, Tugcu T, Chae CB (2014) Three-dimensional channel characteristics for molecular communications with an absorbing receiver. *IEEE Commun Lett* 18(6):929–932

Reception Mechanisms

- [Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion](#)

Reconciliation

- [Wireless Key Establishment](#)

Reduced Guard Interval OFDM/OFDMA

- [Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation](#)

Reinforcement Learning-Based Wireless Communications Against Jamming and Interference

Liang Xiao

Department of Communication Engineering,
Xiamen University, Xiamen, China

Definition

Learning-based anti-jamming communication strategy applies reinforcement learning algorithms for mobile users in wireless networks to achieve the optimal transmission policy against jamming and interference, without knowing the network model, radio channel model, and jamming model.

Historical Background

Due to the broadcast nature of radio propagation, wireless networks are vulnerable to jamming attacks, as jammers purposefully inject replayed

or faked signals into wireless media to interrupt the ongoing radio transmissions between legitimate users (Xu et al. 2005; Xiao 2015). With the pervasion of smart and programmable radio devices such as universal software radio peripherals (USRPs) (Rahbari et al. 2016), smart jammers choose to launch multiple types of attacks, such as eavesdropping and spoofing attacks, and select the jamming power, frequency, and time against the ongoing wireless transmissions (Trappe 2015; Xiao et al. 2018a). Smart jammers can even analyze the ongoing anti-jamming transmission policy and induce the mobile devices to use a specific communication mode and then block them accordingly (Akyildiz et al. 2008; Xiao et al. 2015). Jamming attackers aim to degrade the communication efficiency of the ongoing transmission, increase power consumption of the radio nodes, and even lead to denial of services (DoS) attacks (Xiao et al. 2012a).

Traditional anti-jamming wireless communication solutions, spread spectrum techniques, such as frequency hopping and direct-sequence spread spectrum, have been used for decades to address jamming attacks in wireless networks (Xu et al. 2005; Wang et al. 2012; Xiao et al. 2013). However, significant challenges have to be addressed in 5G systems, e.g., spread spectrum technique requires both the transmitter and the receiver to share the physical-layer secrets such as the spreading codes or frequency hopping pattern in advance. However, jammers can derive the spreading codes of the transmitter by eavesdropping the public control channels and compromising cognitive radio nodes in large-scale dynamic wireless network (Attar et al. 2012).

Anti-jamming techniques such as uncoordinated frequency hopping (Liu et al. 2010; Xiao et al. 2012b) have been proposed to address these problems. However, many of these still suffer from low communication efficiency and high-energy cost in dynamic wireless networks (Li et al. 2010, 2012; Xiao et al. 2012c). In wireless communications, a mobile device has difficulty obtaining the network and jamming model, such as the channel condition, jamming, and interference strength in dynamic environments.

The transmission policy such as the transmit power and channel selection in a dynamic game among the mobile devices, jammers, and interference sources can be viewed as a Markov decision process (MDP) (Xiao et al. 2017, 2018b,c). Reinforcement learning (RL) techniques, such as Q-learning and DQN, can be used to find the optimal strategy in MDP via trial and error (Sutton and Barto 1998). Therefore, the channel selection based on Q-learning can improve the resistance against jamming (Conley and Miller 2013), and the channel accessing with Q-learning can address smart jammers in hostile environments (Gwon et al. 2013). Deep reinforcement learning technique, such as DQN, can accelerate the learning speed of mobile device to achieve the optimal anti-jamming strategy, in the environment with a large number of frequency channels (Han et al. 2017).

Foundations

We assume a wireless communication system using frequency hopping over N frequency channels. At time slot k , the mobile device chooses its transmission policy denoted by $x^{(k)} \in \{0, \dots, N\}$ and sends a message with power P_s on channel $x^{(k)}$. If $x^{(k)} = 0$, the mobile device leaves the area for better transmission condition. The mobile device observes the state at time k denoted by $\mathbf{s}^{(k)}$, which consists of the presence of the PUs, and the previous SINR, i.e., $\mathbf{s}^{(k)} = [\lambda^{(k-1)}, \text{SINR}^{(k-1)}]$.

The CNN is used to estimate the Q-value, denoted by $Q(\mathbf{s}^{(k)}, x)_{0 \leq x \leq N}$, for each action. The CNN consists of two convolutional (Conv) layers and two fully connected (FC) layers. The first Conv layer includes 20 filters each with size 3×3 and stride 1, and the second Conv layer has 40 filters each with size 2×2 and stride 1. Both Conv layers use the rectified linear unit (ReLU) as the activation function. The first FC layer involves 180 rectified linear units, while the second FC layer has $N + 1$ units for the action set. The filter weights of the four layers in the CNN at time k are denoted by $\theta^{(k)}$.

Let $\varphi^{(k)}$ denote the state sequence at time k , which consists of the current system state and the previous W system state-action pairs, i.e., $\varphi^{(k)} = [\mathbf{s}^{(k-W)}, x^{(k-W)}, \dots, x^{(k-1)}, \mathbf{s}^{(k)}]$. The state sequence is then reshaped into a 6×6 matrix as the input to the CNN to estimate $Q(\varphi^{(k)}, x|\theta^{(k)})$, $\forall 0 \leq x \leq N$. The CNN parameters $\theta^{(k)}$ are updated at each time slot based on experience replay.

The experience observed by the mobile device is denoted by $\mathbf{e}^{(k)} = [\varphi^{(k)}, x^{(k)}, u_s^{(k)}, \varphi^{(k+1)}]$, and the memory pool at time k is given by $\mathcal{D} = \{\mathbf{e}^{(1)}, \dots, \mathbf{e}^{(k)}\}$. The experience replay chooses an experience $\mathbf{e}^{(d)}$ from memory pool $\mathcal{D}$ at random, with $1 \leq d \leq k$ to update $\theta^{(k)}$ according to a stochastic gradient descent (SGD) algorithm. The mean-squared error of the target optimal Q-function value is minimized with minibatch updates, and the loss function is chosen by Mnih et al. (2015) as

$$L(\theta^{(k)}) = \mathbb{E}_{\varphi^{(k)}, x, u_s, \varphi^{(k+1)}} \left[\left(R - Q(\varphi^{(k)}, x; \theta^{(k)}) \right)^2 \right], \quad (1)$$

where R is the target optimal Q-function, which is given by

$$R = u_s + \gamma \max_{x'} Q(\varphi^{(k+1)}, x'; \theta^{(k-1)}). \quad (2)$$

The gradient of the loss function with respect to the weights $\theta^{(k)}$ is given by

$$\begin{aligned} \nabla_{\theta^{(k)}} L(\theta^{(k)}) &= \mathbb{E}_{\varphi^{(k)}, x, u_s, \varphi^{(k+1)}} \left[R \nabla_{\theta^{(k)}} Q(\varphi^{(k)}, x; \theta^{(k)}) \right] \\ &\quad - \mathbb{E}_{\varphi^{(k)}, x} \left[Q(\varphi^{(k)}, x; \theta^{(k)}) \nabla_{\theta^{(k)}} Q(\varphi^{(k)}, x; \theta^{(k)}) \right]. \end{aligned} \quad (3)$$

This process repeats B times at each time slot, and $\theta^{(k)}$ is updated according to the B randomly selected experiences.

By using the ε -greedy algorithm, the mobile device chooses the action $x^{(k)}$ according to the

Q-function and $\mathbf{s}^{(k)}$. If $x^{(k)}$ is 0, the mobile device leaves the area. Otherwise, the mobile device transmits the message on channel $x^{(k)}$. The BS estimates the SINR of the device signal and feeds it back to the mobile device. The mobile device evaluates the utility $u_s^{(k)}$ based on the SINR and the transmission cost. According to the next state sequence $\varphi^{(k+1)}$, the mobile device stores the new experience $[\varphi^{(k)}, x^{(k)}, u_s^{(k)}, \varphi^{(k+1)}]$ in the memory pool $\mathcal{D}$.

Key Application

The RL-based anti-jamming wireless communication scheme can be applied in mobile communications and Internet of things.

Cross-References

- [Artificial Intelligence](#)
- [Game Theory](#)
- [Interference Management](#)
- [Wireless Data Security](#)

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant 61671396.

References

- Akyildiz IF, Lee WY, Vuran MC, Mohanty S (2008) A survey on spectrum management in cognitive radio networks. *IEEE Commun Mag* 46(4):40–49
- Attar A, Tang H, Vasilakos AV, Yu FR, Leung VCM (2012) A survey of security challenges in cognitive radio networks: solutions and future research directions. *Proc IEEE* 100(12):3172–3186
- Conley WG, Miller AJ (2013) Cognitive jamming game for dynamically countering ad hoc cognitive radio networks. In: *Military communications conference, MILCOM 2013-2013 IEEE*. IEEE, San Diego, CA, pp 1176–1182
- Gwon Y, Dastangoo S, Fossa C, Kung H (2013) Competing mobile network game: embracing antijamming and jamming strategies with reinforcement learning. In: *2013 IEEE conference on communications and network security (CNS)*. IEEE, National Harbor, MD, pp 28–36

- Han G, Xiao L, Poor HV (2017) Two-dimensional anti-jamming communication based on deep reinforcement learning. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), New Orleans, LA, pp 1–5
- Li C, Dai H, Xiao L, Ning P (2010) Analysis and optimization on jamming-resistant collaborative broadcast in large-scale networks. In: 2010 conference record of the forty fourth asilomar conference on signals, systems and computers (ASILOMAR). IEEE, Pacific Grove, CA, pp 1859–1863
- Li C, Dai H, Xiao L, Ning P (2012) Communication efficiency of anti-jamming broadcast in large-scale multi-channel wireless networks. *IEEE Trans Signal Process* 60(10):5281–5292
- Liu A, Ning P, Dai H, Liu Y (2010) USD-FH: jamming-resistant wireless communication using frequency hopping with uncoordinated seed disclosure. In: 2010 IEEE 7th international conference on mobile adhoc and sensor systems (MASS). IEEE, San Francisco, CA, pp 41–50
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fiedjeland AK, Ostrovski G (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533
- Rahbari H, Krunz M, Lazos L (2016) Swift jamming attack on frequency offset estimation: the achilles heel of OFDM systems. *IEEE Trans Mob Comput* 15(5):1264–1278
- Sutton R, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge
- Trappe W (2015) The challenges facing physical layer security. *IEEE Commun Mag* 53(6):16–20
- Wang Q, Xu P, Ren K, Li XY (2012) Towards optimal adaptive UFH-based anti-jamming wireless communication. *IEEE J Sel Areas Commun* 30(1):16–30
- Xiao L (2015) Anti-jamming transmissions in cognitive radio networks. Springer, Cham
- Xiao L, Chen Y, Lin WS, Liu KR (2012a) Indirect reciprocity security game for large-scale wireless networks. *IEEE Trans Inf Forensics Secur* 7(4):1368–1380
- Xiao L, Dai H, Ning P (2012b) Jamming-resistant collaborative broadcast using uncoordinated frequency hopping. *IEEE Trans Inf Forensics Secur* 7(1):297–309
- Xiao L, Dai H, Ning P (2012c) MAC design of uncoordinated FH-based collaborative broadcast. *IEEE Wirel Commun Lett* 1(3):261–264
- Xiao L, Yan Q, Lou W, Chen G, Hou YT (2013) Proximity-based security techniques for mobile users in wireless networks. *IEEE Trans Inf Forensics Secur* 8(12):2089–2100
- Xiao L, Chen T, Liu J, Dai H (2015) Anti-jamming transmission Stackelberg game with observation errors. *IEEE Commun Lett* 19(6):949–952
- Xiao L, Li Y, Dai C, Dai H, Poor HV (2017) Reinforcement learning-based NOMA power allocation in the presence of smart jamming. *IEEE Trans Veh Technol* 67(4):3377–3389

- Xiao L, Jiang D, Wan X, Su W, Tang Y (2018a) Anti-jamming underwater transmission with mobility and learning. *IEEE Commun Lett* 22:542–545
- Xiao L, Lu X, Xu D, Tang Y, Wang L, Zhuang W (2018b) UAV relay in VANETs against smart jamming with reinforcement learning. *IEEE Trans Veh Technol* 67:4087–4097
- Xiao L, Wan X, Dai C, Du X, Chen X, Guizani M (2018c) Security in mobile edge caching with reinforcement learning. *IEEE Wirel Commun Mag* 25(3):116–122
- Xu W, Trappe W, Zhang Y, Wood T (2005) The feasibility of launching and detecting jamming attacks in wireless networks. In: Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing. ACM, New York, NY, pp 46–57

Relative Localization

- [Cooperative Localization in Wireless Sensor Networks](#)

Relay-Aided Device-to-Device (D2D) Communications

- [Relay-Involved Device-to-Device Communications in LTE Networks](#)

Relay-Assisted D2D Communications

- [Relay-Involved Device-to-Device Communications in LTE Networks](#)

Relaying in LTE-Advanced

Yi Lou¹ and Honglin Zhao²

¹College of Underwater Acoustic Engineering, Harbin Engineering University, Harbin, China

²Department of Communication Engineering, Harbin Institute of Technology, Harbin, China

Synonyms

[Relays in LTE-Advanced](#)

Definition

Relaying and many other technologies have been standardized by Third-Generation Partnership Project (3GPP) in Release 10 of Long-Term Evolution (LTE) specifications, which results in huge performance enhancements compared to Release 9. To underline the major evolution, 3GPP approved LTE-Advanced as the official marker for Release 10 of LTE. In LTE-Advanced, relaying means that the eNodeB communicates with the user equipment (UE) with the help of a relay node (RN) wirelessly connected to the network. Specifically, the RN receives the RF signals transmitted from the eNodeB/UE and then forwards the processed signals to the UE/eNodeB.

Historical Background

With the increasing demand for high-rate applications and explosive growth in the number of UEs, the performance requirements of the wireless network, such as capacity and data rates, are much higher. One way of network performance enhancement is to deploy more eNodeBs. In this way, the distance between base stations and the distance between the eNodeB and the corresponding served UEs are both shortened which leads to higher data rates for each UE. However, to connect to the network, an eNodeB requires wired backhaul link such as an optical fiber, which results in a high-cost deployment and long construction period. Moreover, in some places, such as rural areas, deploying eNodeBs is not cost-effective as of which the functionality cannot be fully exploited due to the low population density. In order to alleviate the above problems in the deployment aspect, 3GPP considered relaying as one of the critical technologies of LTE-Advanced. 3GPP started relaying work item since January 2010 and standardized the corresponding specification at April 2011 (3GPP 2011a,b). Relaying work item was started from January 2010 and formally standardized in March 2011.

Relays can be classified in terms of a variety of criteria. According to the processing method for the received signals, relays can be classified as amplify-and-forward (AF) and decode-and-forward (DF) ones. AF relays amplify the received signals in radio frequency and then forward them to the destination node. DF relays decode and re-encode the received signals before forwarding them to the destination node. Note that, the noise and interference are also amplified along with the effective information components by using AF relays, which are eliminated by using the DF ones. According to the operating bands used by the transmitter and the receiver, relays can be classified as out-band and in-band ones. For the out-band relays, the transmitter and the receiver in the relay operate on different frequency bands. The main advantage of out-band relays is to support full-duplex operation, i.e., out-band relays can simultaneously transmit and receive signals, as long as the separation in the frequency domain of the transmission and reception links is large enough so that there is no interference with each other. Moreover, out-band relays can be integrated into the network without any modification of the radio interfaces of the previous release eNodeBs and UEs. However, the costs of network operation are increased due to the additional required spectrum. For the in-band relays, the transmission and reception links operate in the same frequency band. For this case, time division multiplexing should be used for the two links to make sure that only one link is active at a given time, i.e., half-duplex, thereby avoiding interference with each other. It should be noted that full-duplex operation can still be achieved for the in-band relays by physically isolating between the antennas of the transmitter and receiver in in-band relays. However, the overall costs will be increased seriously. By August 2009, 3GPP identified two types of RN, i.e., type-1 RNs and type-2 RNs during the research term of LTE-Advanced relaying. Both types of RN are DF relays. The main difference between the two types is whether they are transparent to the served UEs. A type-1 RN has its own physical cell ID (PCI, defined in LTE Release 8) and generates a new cell different from the linked donor eNodeB

cell. A type-1 RN independently generates cell control signaling and transmits scheduling information to the served UEs. Therefore, from UEs' point of view, a type-1 RN acts as a separate eNodeB distinct from the donor cell. Also, a type-1 RN should act as if it is a Release 8/9 eNodeB to the Release 8/9 UEs with the aim of backward compatibility. A type-2 RN cannot generate its own PCI and therefore share the same PCI used by the donor eNodeB. A type-2 RN is used only for the physical downlink shared channel (PDSCH) transmission since it cannot generate cell control signaling and does not have its own scheduler. The primary application of type-2 RNs is the cooperative relaying (Dohler and Li 2010).

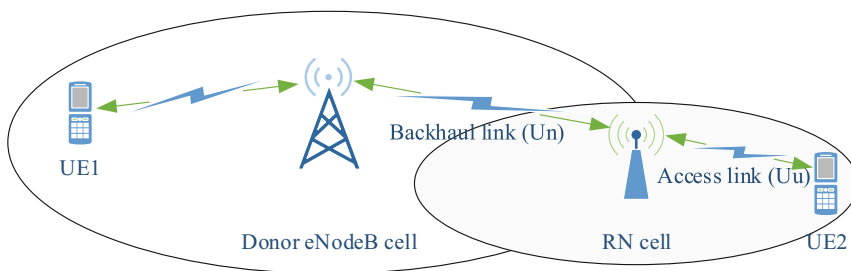
In LTE-Advanced, the relaying standardization mainly focuses on the in-band half-duplex type-1 RNs. In the context of LTE-Advanced and from a functional perspective, an RN can be seen as an eNodeB with the functionality of wireless backhauling. Explicitly, RNs wirelessly connect to the network using LTE radio-access interface technologies which allow RNs to be deployed quickly and with lower costs. However, the wireless backhauling also brings some disadvantages to RNs compared to eNodeBs. The overall delay between the eNodeB and the UEs served by an RN is increased due to the wireless backhaul propagation. Meanwhile, since the access resource is shared between the backhaul link and access link – the link between the RN and the UE served by the RN, the RN can use only partial access resource to avoid self-interference and therefore offers a smaller capacity than one eNodeB.

A typical RN deployment scenario is shown in Fig. 1. The donor eNodeB can only provide

services to the UE located in the donor eNodeB cell, i.e., UE1. To offer services to UE2 located out of the donor eNodeB cell, an RN can be deployed by the network operator at a place in the donor eNodeB cell so that UE2 is located in the RN cell. An RN communicates with the donor eNodeB via backhaul link, which is also referred to as Un interface (3GPP 2015), and communicates with a UE via access link which is also referred to as Uu interface. The term “donor” in front of eNodeB implies that the eNodeB shares partial radio resources belonging to the UEs located in the eNodeB cell to RNs for servicing UEs out of the eNodeB cell. To seamlessly integrate RNs into the implemented LTE networks, 3GPP considered the backward compatibility with the previous release eNodeBs as the primary requirement, which will be further discussed in section “Foundations.”

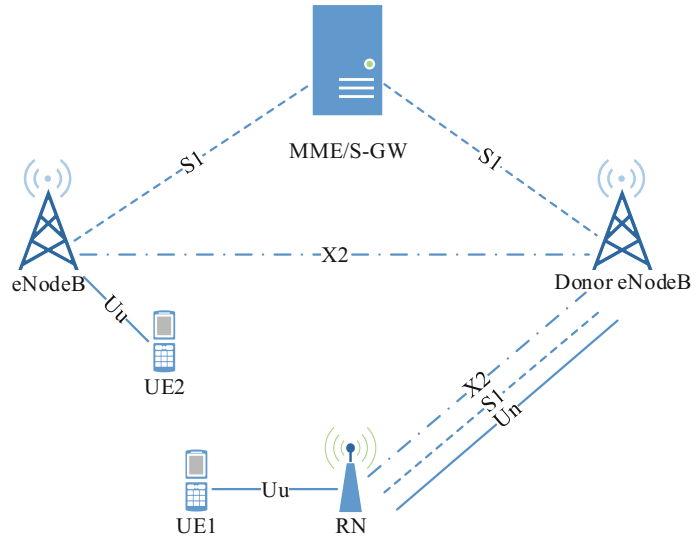
Foundations

The overall architecture of LTE-Advanced relaying is illustrated in Fig. 2. Both eNodeB and donor eNodeB are connected to the mobility management entity (MME) and serving gateway (S-GW) units of the core network via S1 interface. X2 interface provides a connection between eNodeBs. The first requirement of relaying standardization is backward compatibility with Release 8/9 UEs; that is, there is no difference between connecting an RN or an eNodeB from the perspective of UE1. Therefore, an RN has all the functionality of an eNodeB. As shown in Fig. 2, the access link between the RN and UE1 also follows the same Uu interface specification



Relaying in LTE-Advanced, Fig. 1 A typical RN deployment scenario

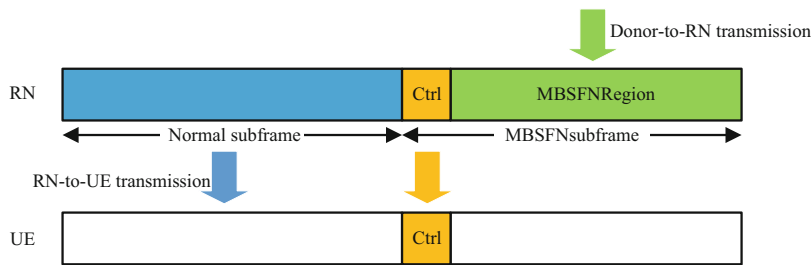
**Relaying in
LTE-Advanced, Fig. 2**
Overall relay architecture



used for the connection between UE2 and the eNodeB. An RN also has the basic functionality of a UE along with relay-specific enhancements to ensure that the RN can communicate with the donor eNodeB via the Un interface. In order to communicate with other nodes of the network such as eNodeBs, MME, and S-GW, an RN needs the S1 and X2 interfaces, which are offered by the donor eNodeB via the Un interface. Therefore, the donor eNodeB acts like a proxy between the served RN and the rest of nodes of the network. From the perspective of the core network, it looks like UE1 is directly connected to the donor eNodeB. From the perspective of the RN, it looks like the RN communicates with the core network directly. Also, for the X2 interface proxy, the donor eNodeB is also transparent as if the RN communicates with the other eNodeB via the X2 interface. The RN acts like the donor eNodeB toward the served UE and acts like the served UE toward the donor eNodeB. Therefore, the RF requirements for the LTE-Advanced RN are similar to those of an eNodeB on the access link and those of a UE on the backhaul link (3GPP 2012, 2017a,b).

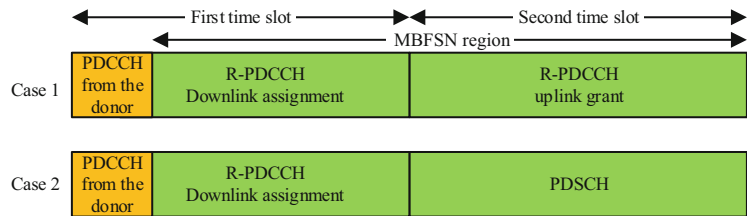
For the in-band relaying scenario, RNs cannot simultaneously transmit and receive signals for avoiding mutual interference. In LTE-Advanced,

RNs exploit time division multiplexing between the transmission of the backhaul and access links. Therefore, the UE-to-RN transmission is not allowed at the subframe that the RN takes for transmitting signals to the donor eNodeB. For this purpose, the uplink scheduler in the RN can be used to ensure that the access link is not activated for such subframe. According to a similar principle, the RN can only receive signals from the eNodeB for the subframe where no RN-to-UE transmission takes place. However, Release 8/9 UEs at least expect the synchronization signals and reference signals (CRS) in each subframe from the access downlink transmission to keep the connection with the RN. To simultaneously keep compatible with the Release 8/9 UEs and receive signals from backhaul downlink, RNs can adopt multicast-broadcast single-frequency network (MBSFN) subframe, which is defined in the previous release of LTE for supporting multimedia broadcast/multicast service. Release 8/9 UEs can recognize the MBSFN subframe where the first or two OFDM symbols are referred to as the control region and the remaining symbols are referred to as the MBSFN region. An RN can then use an MBSFN subframe to transmit at least synchronization signals and CRS to the UES in the control region and then receives the donor-to-RN transmission in the



Relaying in LTE-Advanced, Fig. 3 Time division multiplexing between backhaul and access links

Relaying in LTE-Advanced, Fig. 4 A typical usage of R-PDCCH



MBSFN region, as shown in Fig. 3. The subframe type used by the UE, i.e., normal and MBSFN subframe, is scheduled by the radio resource control signaling.

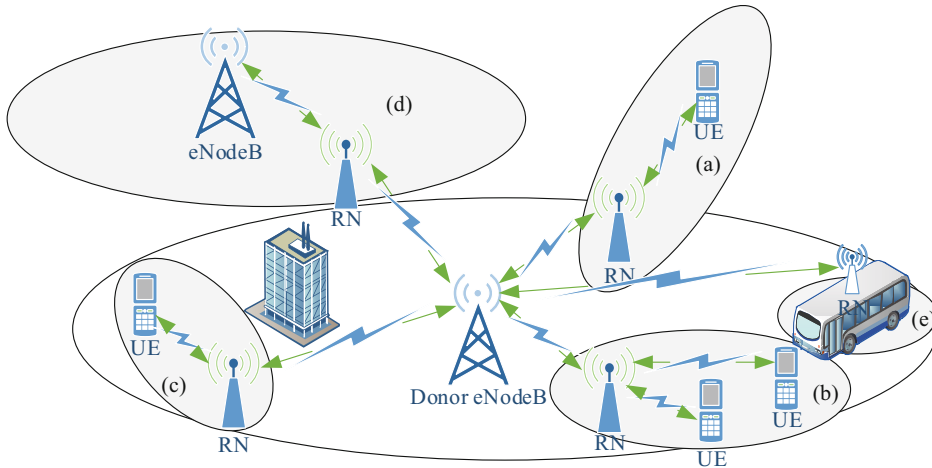
The time-frequency resource blocks (RBs) used for the necessary downlink control information and the RBs used for paging information and data transmission are, respectively, defined as physical downlink control channel (PDCCH) and PDSCH. The RN cannot receive the L1/L2 control signaling transmitted from the eNodeB in the PDCCH using the OFDM symbols in the control region of the MBSFN subframe, which is already occupied by the RN-to-UE synchronization signals and CRS transmission. To cater for this, 3GPP introduced an RN-specific PDCCH, i.e., R-PDCCH in Release 10 of LTE. R-PDCCH is transmitted using the OFDM symbols in the MBSFN region of the MBSFN subframe and therefore should be multiplexed with PDSCH. A typical usage of R-PDCCH is shown in Fig. 4. The downlink assignment to the RN is transmitted in the first time slot of the MBSFN subframe. On the other hand, the low time-priority uplink grant is located in the second time slot since the uplink transmission is later than downlink. When no uplink grant needs to be

transmitted, the second time slot can be utilized for PDSCH transmission from the eNodeB to the RN.

Key Applications

RNs can be used as eNodeBs for coverage extension and capacity enhancement. Also, wireless backhauling makes RNs suitable for many application scenarios because of the advantages of flexible deployment and cost-effective implementation. The typical application scenarios are displayed as follows (Yuan 2012):

- *Rural area.* The required traffic volumes in rural area are usually small due to the low population density. From a network operator's point of view, instead of deploying eNodeBs, deploying RNs for covering such area is a better choice as it can reduce the costs of deployment without degrading the quality of service. In this case, coverage extension is the primary concern as shown in Fig. 5a.
- *Hot spot.* The network operator can deploy RNs in outdoor/indoor hot spots for capacity enhancement to support a satisfactory service



Relaying in LTE-Advanced, Fig. 5 Key application scenarios

experience (throughput) for the UEs in such places, as shown in Fig. 5b. The transmission power of RNs should keep low to reduce interferences to other cells.

- *Dead spot.* In urban areas, UEs may be located in a shadowing area due to high-rise buildings. For this case, RNs can be deployed to eliminate such coverage holes by forming line-of-sight backhaul and access links as shown in Fig. 5c.
- *Wireless backhaul.* For areas with low spectrum utilization, such as mountainous and rural areas, RNs can be deployed for wireless backhauling, i.e., the eNodeB-to-eNodeB communication as shown in Fig. 5d.
- *Temporary coverage.* Due to the characteristics of wireless backhauling, RNs are suitable for temporary radio coverage. Note that, deployment flexibility can be further enhanced by omitting the wired power connection by means of wireless energy harvesting technology (Zhang and Ho 2013).
- *Group mobility.* Due to the Doppler effects, mobile UEs in a bus communicate with the eNodeB via fast-fading channels and therefore consume more power to maintain a satisfactory data rate (Wu and Fan 2016). As shown in Fig. 5e, an RN can be mounted on the roof of the bus to provide static and stable access link for all the onboard UEs and hence save their battery consumption.

Cross-References

- [Relay-Involved Device-to-Device Communications in LTE Networks](#)

References

- 3GPP (2011a) Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2. TS 36.300, 3GPP, version 10.3.1
- 3GPP (2011b) Evolved universal terrestrial radio access (E-UTRA); Physical layer for relaying operation. TS 36.216, 3GPP, version 10.3.1
- 3GPP (2012) Evolved universal terrestrial radio access (E-UTRA); Relay radio transmission and reception. TS 36.116, 3GPP, version 11.0.0
- 3GPP (2015) Evolved universal terrestrial radio access network (E-UTRAN); Architecture description. TS 36.401, 3GPP, version 12.2.0
- 3GPP (2017a) Evolved universal terrestrial radio access (E-UTRA); Base station (BS) radio transmission and reception. TS 36.104, 3GPP, version 14.3.0
- 3GPP (2017b) Evolved universal terrestrial radio access (E-UTRA); User equipment (UE) radio transmission and reception. TS 36.101, 3GPP, version 14.3.0
- Dohler M, Li Y (2010) Cooperative communications: hardware, channel and PHY. Wiley, New York
- Wu J, Fan P (2016) A survey on high mobility wireless communications: challenges, opportunities and solutions. IEEE Access 4:450–476
- Yuan Y (2012) LTE-Advanced relay technology and standardization. Springer, Heidelberg/Berlin
- Zhang R, Ho CK (2013) MIMO broadcasting for simultaneous wireless information and power transfer. IEEE Trans Wirel Commun 12(5):1989–2001

Relay-Involved Device-to-Device Communications in LTE Networks

Ruofei Ma, Bo Li, and Gongliang Liu
Department of Communication Engineering,
Harbin Institute of Technology, Weihai, China

Synonyms

Relay-Aided Device-to-Device (D2D) Communications; Relay-Assisted D2D Communications

Definitions

D2D communication enables direct data transmissions between two D2D-capable user equipments (UEs) bypassing base station (BS) in Long-Term Evolution (LTE) systems. It can also be called direct D2D communication. Note: in the following text, evolved Node B (eNB) is used to replace the expression of BS.

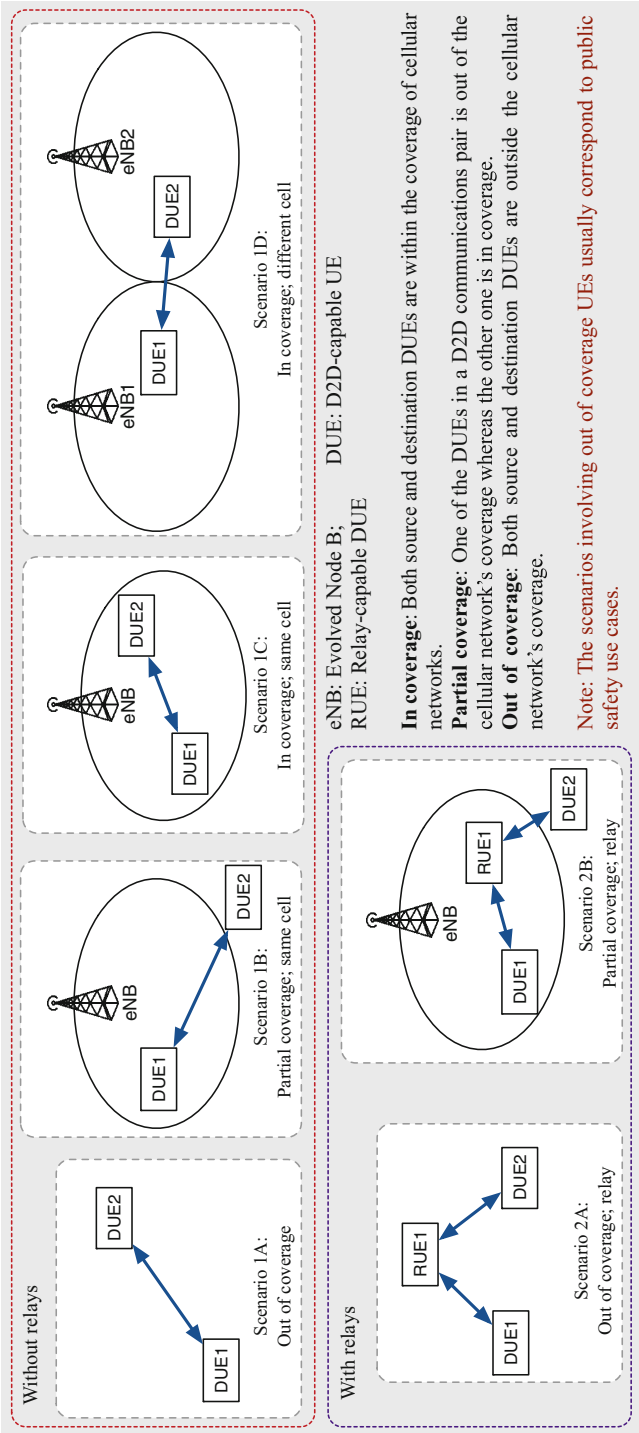
Relay-involved D2D communication encourages two D2D-capable UEs (DUEs) to communicate with each other via a relay node, which can be considered as a broad sense D2D communication. It extends the range of direct D2D communication.

Background

The concept of D2D communication was first proposed in Lin and Hsu (2000) to depict the direct data transmission between neighbor UEs on a multi-hop relaying link in cellular networks. Later on, the potential of using D2D communication to improve spectrum efficiency and system capacity of cellular networks was investigated. In the most typical research by Doppler et al. (2009), D2D communication was first treated as an underlay to LTE-Advanced (LTE-A) networks, in which mechanisms for D2D communication setup and management

involving procedures in the LTE system architecture evolution were proposed, and including D2D communication can increase the cell-wise throughput was proved. Soon after, other potential D2D use cases were identified, such as contents sharing, peer-to-peer data exchanging (Lei et al. 2012), UE-to-eNB data relaying, and so on. The first testbed for integration of D2D communication into a cellular network was the FlashLinQ built by Qualcomm, which is actually a physical (PHY)/media access control (MAC) architecture for D2D communications underlaying cellular networks (Asadi et al. 2014). By the use of orthogonal frequency division multiplexing/orthogonal frequency division multiple access (OFDM/OFDMA) and distributed scheduling, FlashLinQ provides researchers reference solutions for efficient proximity discovery, timing synchronization, and link management in D2D-involved cellular networks. These researches prove that not only can D2D communication improve the entire system spectral efficiency via flexible reuse of radio resources but also ease the core network data processing loads by performing traffic offloading. Besides, the Third Generation Partner Project (3GPP) has also started the discussions on developing D2D communications as proximity services (ProSe) in LTE networks, and Fig. 1 shows the D2D scenarios taken into account by the 3GPP standardization group (Mach et al. 2015).

D2D communication can be classified into licensed-band D2D and unlicensed-band D2D. The unlicensed-band D2D communication is performed based on other wireless network standards operating on unlicensed radio resources, which can support direct communications, such as Wi-Fi Direct. It can be realized either in the network-controlled manner or in the autonomous manner. The licensed-band D2D communication takes place on the licensed cellular radio resources allocated by the core network. Licensed-band D2D communications are all in network-controlled manner and can be further divided into dedicated (i.e., overlay) and reuse (i.e., underlay) modes. The former works in a way that D2D UEs operate on



Relay-Involved Device-to-Device Communications in LTE Networks, Fig. 1 3GPP D2D communication scenarios

separated cellular resources, whereas the latter requires that D2D UEs share the same cellular resources with general cellular UEs (CUEs). In dedicated mode, there is no interference between D2D and cellular UEs, due to the fact that eNB assigned orthogonal channels for D2D and cellular links. In reuse mode, there are co-channel interferences between D2D and the shared cellular links, which means that eNB should keep such interference lower enough to ensure the quality of service (QoS) of both D2D and the shared cellular links when performing joint scheduling on resource allocation and power control. Working on unlicensed bands is commonly subject to uncontrolled interference from coexistence systems. Hence, majority of the existing research focused on licensed-band D2D communications within a single cell environment just as Scenario 1C in Fig. 1, aiming to improve spectrum efficiency and cell-wise data transmission performance by encouraging flexible resource allocation and communication mode selection (Fodor et al. 2012; Min et al. 2011; Yu et al. 2014). These works usually assume three communication modes for D2D-capable UEs, i.e., general cellular mode, dedicated D2D mode, and reuse D2D mode, as shown in Fig. 2. The dedicated mode only utilizes traffic offloading capability of D2D communications, whereas the reuse mode could further obtain channel reuse gain. Note that it is suggested that D2D UEs in reuse mode should reuse the cellular uplink resources to avoid the interference from eNB.

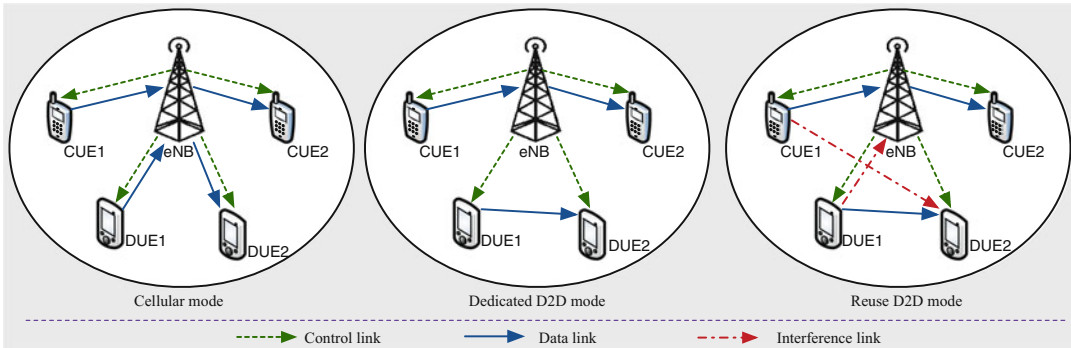
Both dedicated and reuse D2D modes in existing works actually belong to direct D2D mode, i.e., the source and destination DUEs perform direct communication with each other. Only using direct D2D mode may limit the benefits brought in by D2D communications to LTE systems, because sometimes source and destination DUEs may not be able to perform direct D2D communications due to long separation distance and poor channel condition between them (Hasan et al. 2014). In such cases, a promising way to fully explore the benefits of D2D communications is to enable eNB-controlled D2D transmissions through relays,

which can substantially extend the range of D2D communications (Ni 2016).

Bringing relay technologies into D2D communications offers DUEs more opportunities to work on D2D mode in a cellular network, and it can be regarded as a supplement/enhancement to direct D2D. Currently, the research in this area is still in its initial exploratory stage. This work tried to offer audiences a short overview on the current research on relay-involved D2D communications.

Potential Relay-Involved D2D Communications

In existing research focusing on the D2D communication with relay assistance, the relay node can be either relay-capable UEs (RUEs) or fixed relay stations. Hasan et al. (2014) investigated D2D communications served by fixed location LTE-A Layer-3 (L3) relay stations, which can relay not only data from faraway UEs to eNB but also data between source and destination DUEs. They mainly discussed the radio resource and power allocation problem at the relay node when the data from UE to eNB and from source DUE to destination DUE are simultaneously relayed. eNB can also be treated as a relay. Ma et al. (2017b) proposed local route mode as another D2D communication mode, in which the source and destination DUEs communicate with each other using eNB as their relay station, but the data will not go through the core network. Here eNB acts as only a relay station with scheduling capability. But RUE-assisted D2D communications may be more flexible due to an explosive increase in UE density in LTE systems. Hence most works in this area considered the “relays” in relay-involved D2D communication as RUEs. Some literatures called relay-involved D2D communication as two-hop D2D, which can be further classified into two types. The first type demonstrates that a DUE in/out of a eNB’s coverage communicates with the eNB via relay-capable DUE’s relaying with an aim to enhance the transmissions to/from eNB or extend the coverage of eNB (Nishiyama et al. 2014; Liu and Kato



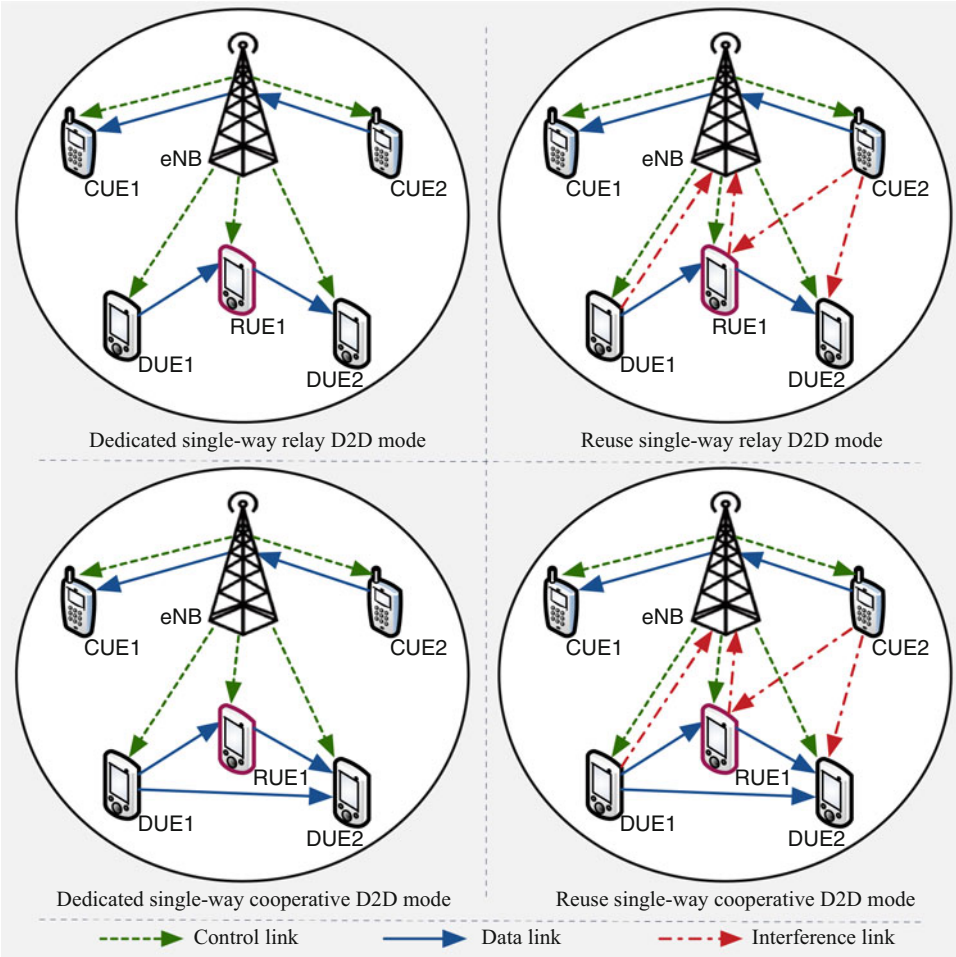
Relay-Involved Device-to-Device Communications in LTE Networks, Fig. 2 Three typical communication modes for DUEs in existing literatures

2015), whereas the second type demonstrates that two DUEs within an eNB's coverage may communicate with each other via the help of a relay-capable DUE under the eNB's control when direct D2D cannot meet the QoS requirements (Wei et al. 2016; Ma et al. 2012). Actually, in the first type two-hop D2D, there is only one hop transmission (i.e., DUE to/from relay) on the DUE-relay-eNB link strictly belonging to D2D communications, and the other hop transmission (i.e., Relay to/from eNB) is just same as the cellular uplink or downlink communication. In strict sense, it cannot be called two-hop D2D. Thus, this work focused on introduction of two-hop D2D belonging to the second type, in which both source and destination nodes are DUEs.

According to different relaying schemes adopted, the research on relay-involved D2D communications can be highlighted and summarized from the following four directions:

- *Single-way relay-assisted D2D communication*: It demonstrates that two DUEs communicate with each other with the help of a selected RUE, under which a complete single-way end-to-end data transmission is finished in two time slots (Ma et al. 2012, 2017a). The source-to-relay and relay-to-destination transmission are sequentially performed. The selected RUE receives the signals transmitted from source DUE in the first time slot and makes necessary processing, and then it transmits the processed

signals to destination DUE. eNB can allocate the source-to-relay and relay-to-destination links either the same radio resources or the different radio resources. Relay-assisted D2D communication can also be classified into dedicated mode and reuse mode according to whether it is permitted to share the CUEs' channels. Figure 3 shows four potential modes for single-way relay-assisted D2D communications, i.e., dedicated/reuse single-way relay D2D and dedicated/reuse single-way cooperative D2D. In the reuse single-way relay D2D mode of Fig. 3, the source-to-relay and relay-to-destination D2D links reuse the same cellular uplink channel, i.e., CUE2's uplink channel. Hence, in the first hop transmission, the receiving relay node, RUE1, suffers interference from CUE2, and the CUE2's uplink transmission suffers interference from the source DUE, i.e., DUE1, whereas in the second hop transmission, the interference links are just from CUE2 to the receiving DUE2 and from the transmitting RUE1 to the receiving eNB. The cooperative D2D mode can be considered as a further enhancement to the corresponding relay D2D mode, under which the source DUE transmits desired signals to the relay node and destination DUE simultaneously in a broadcasting manner in first transmission time slot and, in the second transmission time slot, the relay node will also transmit the received signals to the destination DUE. Here, after



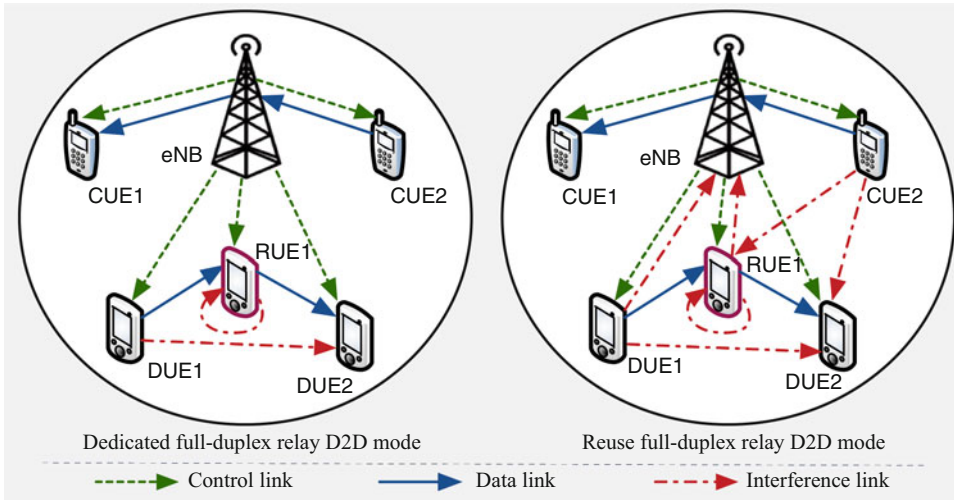
Relay-Involved Device-to-Device Communications in LTE Networks, Fig. 3 Single-way relay-assisted D2D communication modes

R

receiving both signals from the relay link and direct link, destination DUE will make a combination processing to them to strengthen the desired signal. Typical combining methods include maximum ratio combining and equal gain combining.

- *Full-duplex relay-assisted D2D communication:* It demonstrates that in the source-relay-destination D2D link, the selected RUE works in a full-duplex manner, meaning that the relay node is capable of performing signal receiving and transmitting at the same time (Wang et al. 2015). Thus, the data transmission from source to destination DUEs via a full-duplex RUE can be finished within only

one time slot, i.e., the source-to-relay and relay-to-destination transmissions take place simultaneously. In such full-duplex relaying scenarios, the source-to-relay and relay-to-destination links are commonly assigned the same radio resources, and this leads to new interferences, including self-interference at the relay node and the interference from source DUE to destination DUE, as depicted in Fig. 4, which shows the potential dedicated and reuse full-duplex relay-assisted D2D communication modes. In the reuse full-duplex relay D2D mode of Fig. 4, the uplink channel of CUE2 is reused by entire relay link, and all the potential interferences occur



Relay-Involved Device-to-Device Communications in LTE Networks, Fig. 4 Full-duplex relay-assisted D2D communication modes

in the same time slot. Apparently, to ensure the QoS requirements for both D2D and cellular links in full-duplex relaying scenarios, eNB needs to be equipped with more complicated scheduling schemes for efficient interference control.

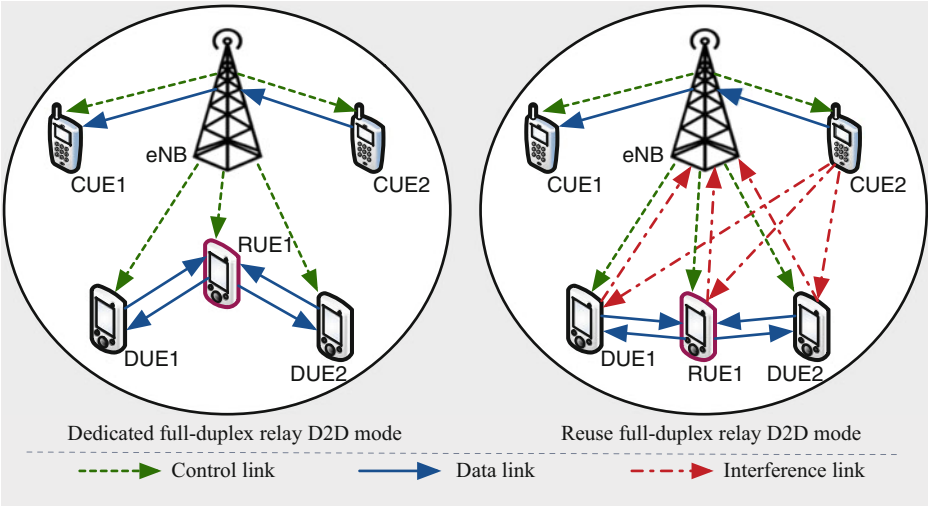
- *Two-way relay-assisted D2D communication:* It demonstrates that in the relay-assisted D2D path, the selected RUE simultaneously receives the signals from source and destination DUEs just like a multiple access process and directly makes a summation to the two signals in the first transmission time slot, and then in the second transmission time slot, the selected RUE broadcasts the combined signal to the source and the destination DUEs (Wei et al. 2016; Ni et al. 2014; Ni 2016). Note that the source and the destination DUEs should transmit their data simultaneously to the relay, and before broadcasting, the relay node may scale the combined signal to ensure the QoS requirements. Once the source and destination DUEs receive the broadcasted signal from the relay node, each of them can acquire its desired term by canceling the self-interference term, since they know their transmitted signals, respectively. Hence, with this relaying scheme, a complete data

exchange between the source and destination DUEs can be finished within only two time slots. According to whether reusing the CUEs' channels, two-way relay-assisted D2D communication can also be classified into dedicated and reuse modes, as shown in Fig. 5. In dedicated mode, same radio resources are assigned to all the transmission links in the two time slots. In reuse mode, the source-to-relay and the destination-to-relay links in the first time slot are assigned the same radio resources, but the radio resources used in the first and the second time slots can be either the same or different. The reuse mode in Fig. 5 shows audiences the situation that the transmissions in the first and the second time slots are reusing the same cellular uplink resources. The interference graph under the reuse mode is clear: in the first time slot, the relay node will receive interference from the CUE whose uplink channel is reused and eNB will simultaneously receive the interferences from both the source and destination DUEs, while in the second time slot, both the source and destination DUEs will be interfered by the CUE whose uplink channel is reused and the eNB will be interfered by the relay node. It should be stated that to realize such

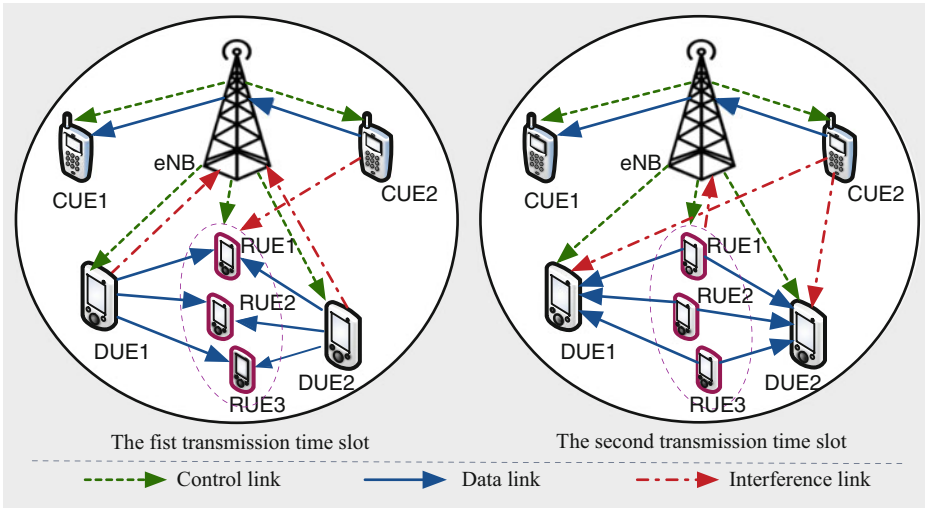
two-way relay-assisted D2D transmissions, more efficient and precise timing management schemes should be designed for eNB, which definitely will increase the system complexity.

- *Network coding (NC)-based relay-assisted D2D communication:* It usually demonstrates the scenarios that multiple D2D communication pairs share the same relay node (Zhao et al. 2017) or one D2D communication

pair performs data exchange using multiple relay nodes at the same time (Huang et al. 2017). Taking the scenario that one D2D communication pair with multiple relay nodes for an example, its data exchanging process, when reusing cellular uplink resources, can be depicted as Fig. 6, in which all the transmission links in the two time slots reuse the same CUE’s uplink channel. Note that radio resource allocated for the transmission



Relay-Involved Device-to-Device Communications in LTE Networks, Fig. 5 Two-way relay-assisted D2D communication modes



Relay-Involved Device-to-Device Communications in LTE Networks, Fig. 6 NC-based relay-assisted D2D communication in reuse mode

R

links in the first time slot can be different from that in the second time slot. It depends on the eNB's scheduling. NC-based relay-assisted D2D communication can also be classified into dedicated and reuse mode according to whether performing channel reuse. Only the reuse mode is introduced here. As shown in Fig. 6, in the first transmission time slot, the source and destination DUEs simultaneously broadcast their signals to all the selected RUEs, and each relay will encode its received data packets. One of the most simple encoding methods is encoding the data packets via XOR operation. The amount of packets encoded at a relay node is just the number of data packets received from DUE1 and DUE2. This is exactly the coding process in the so-called "NC". Due to reuse of the uplink channel of CUE2 by D2D links, DUE1 and DUE2 will cause interference to the cellular uplink, and the relay node will also receive interference from CUE2. In the second transmission time slot, all relay nodes broadcast their encoded data packets to DUE1 and DUE2, which will then perform packets decoding by using their transmitted data packets in the first time slot. The amount of decoded data packets of DUE1, with respect to only one relay node, is the minimum between the packet transmission rate on this relay to DUE1 link and the amount of this relay node's encoded packets in the first time slot. The total amount of decoded data packets of DUE1, respecting to multiple relay nodes, is the sum of the amounts of DUE1's decoded data packets with respect to each relay node. The amounts of DUE2's decoded data packets can be understood in a similar way. The corresponding interferences include the inferences from relay nodes to eNB and from CUE2 to both DUE1 and DUE2. Obviously, it requires new design for the eNB's scheduling to support such a complicated communication process. In addition, it can be easily identified that the aforementioned two-way relay-assisted D2D communication can be considered as a one-relay case of the NC-based relay-assisted D2D communication.

Overall, exiting research on the four directions of relay-assisted communications is still in the stage of performance exploratory. These relay-assisted D2D communication modes have not been taken into account as candidate D2D modes in the mode selection procedure, and they were usually researched from the perspective of end-to-end transmission performance analysis. Furthermore, comparing the four types of relay-assisted D2D communication, it can be easily concluded that full-duplex, two-way, and NC-based relay-assisted D2D modes correspond to more complicated interference control and timing management, which will further increase the degree of difficulty on joint scheduling scheme design at the eNB side. In this respect, single-way relay-assisted D2D modes may be more implementable.

Key Applications

Typical use cases for relay-involved D2D communications may include traffic offloading, contents sharing, local voice service, machine-to-machine (M2M) communication, vehicle-to-vehicle (V2V) communication, and so on.

Key Research Challenges

Due to an explosive increase in both types and amounts of wireless devices, it is still a widely open issue to realize the relay-involved D2D communication modes in LTE networks, and the current LTE standard does not support them. Specifically, to support each of the aforementioned relay-assisted D2D modes raises up a lot of new issues on protocol-level design of an LTE network. The challenges come in three aspects.

- New radio protocol architectures (RPAs) should be designed for both DUEs and RUEs to support relay-assisted D2D modes. Perhaps, RPA design for direct D2D mode can refer to the existing PC5 interface for Proximity Service (ProSe) direct communications in

LTE-A. However, none of the existing releases of the LTE-A standard provided a definition level reference for the RPA design in relay-involved D2D modes.

- An efficient link-quality measurement and report strategy is needed to help the eNB performing control and scheduling on DUEs and RUEs. The involvement of relay-assisted D2D modes makes link-quality assessment and report procedures even more complex. Thus, how to design a precise and low signalling overhead strategy on link-quality measurement and report is critical.
- Scheduling algorithm at eNB should be redesigned to support communication mode selection for all DUEs. In current LTE system designs, the eNB needs only to perform scheduling on radio resource allocation and power control for all its UEs in each subframe. However, if all the related D2D modes are supported by most of the UEs in the next release of LTE-A standard, it will require that the eNB should be able to perform joint scheduling on communication mode selection (including relay selection for the relay-assisted D2D modes), resource allocation, and power control for all DUEs in each subframe to ensure an acceptable system performance.

Cross-References

- [Relaying in LTE-Advanced](#)
- [Resource Management in Macrocell–Small Cell Systems and D2D-Assisted Cellular Systems](#)
- [Sidelink Transmissions and Proximity Services in LTE-A and Beyond](#)
- [WiFi Offloading in WiFi-Cellular Networks](#)

References

- Asadi A, Wang Q, Mancuso V (2014) A survey on device-to-device communication in cellular networks. *IEEE Commun Surv Tutor* 16:1801–1819
- Doppler K, Rinne M, Wijting C, Ribeiro C, Hugl K (2009) Device-to-device communication as an underlay to LTE-advanced networks. *IEEE Commun Mag* 47:42–49
- Fodor G, Dahlman E, Mildh G, Parkvall S, Reider N, Miklos G, Turanyi Z (2012) Design aspects of network assisted device-to-device communications. *IEEE Commun Mag* 50:170–177
- Hasan M, Hossain E, Kim DI (2014) Resource allocation under channel uncertainties for relay-aided device-to-device communication underlying LTE-A cellular networks. *IEEE Trans Wirel Commun* 13:2322–2338
- Huang J, Gharavi H, Yan H, Xing CC (2017) Network coding in relay-based device-to-device communications. *IEEE Netw* 31:102–107
- Lei L, Zhong Z, Lin C, Shen X (2012) Operator controlled device-to-device communications in LTE-advanced networks. *IEEE Wirel Commun* 19:96–104
- Lin YD, Hsu YC (2000) Multihop cellular: a new architecture for wireless communications. In: *IEEE INFOCOM*, vol 3, pp 1273–1282
- Liu J, Kato N (2015) Device-to-device communication overlaying two-hop multi-channel uplink cellular networks. In: *The ACM international symposium on mobile ad hoc networking and computing*, pp 307–316
- Ma X, Yin R, Yu G, Zhang Z (2012) A distributed relay selection method for relay assisted device-to-device communication system. In: *IEEE international symposium on PIMRC*, pp 1020–1024
- Ma R, Chang YJ, Chen HH, Chiu CY (2017a) On relay selection schemes for relay-assisted D2D communications in LTE-A systems. *IEEE Trans Veh Technol* 66:8303–8314
- Ma R, Xia N, Chen HH, Chiu CY, Yang CS (2017b) Mode selection, radio resource allocation, and power coordination in D2D communications. *IEEE Wirel Commun* 24:112–121
- Mach P, Becvar Z, Vanek T (2015) In-band device-to-device communication in OFDMA cellular networks: a survey and challenges. *IEEE Commun Surv Tutor* 17:1885–1922
- Min H, Seo W, Lee J, Park S, Hong D (2011) Reliability improvement using receive mode selection in the device-to-device uplink period underlying cellular networks. *IEEE Trans Wirel Commun* 10:413–418
- Ni Y (2016) Performance analysis of device-to-device communication assisted by mobile relaying. PhD thesis, Nanjing University of Posts and Telecommunications, Nanjing
- Ni Y, Jin S, Wong KK, Zhu H, Shao S (2014) Outage performances for device-to-device communication assisted by two-way amplify-and-forward relay protocol. In: *IEEE WCNC*, pp 502–507
- Nishiyama H, Ito M, Kato N (2014) Relay-by-smartphone: realizing multihop device-to-device communications. *IEEE Commun Mag* 52:56–65
- Wang L, Tian F, Syensson T, Feng D, Song M, Li S (2015) Exploiting full duplex for device-to-device communications in heterogeneous networks. *IEEE Commun Mag* 53:146–152
- Wei L, Hu RQ, Qian Y, Wu G (2016) Energy-efficiency and spectrum-efficiency of multi-hop device-to-device

communications underlaying cellular networks. *IEEE Trans Veh Technol* 65:367–380

Yu G, Xu L, Feng D, Yin R, Li GY, Jiang Y (2014) Joint mode selection and resource allocation for device-to-device communications. *IEEE Trans Commun* 62:3814–3824

Zhao Y, Li Y, Wu D, Ge N (2017) Overlapping coalition formation game for resource allocation in network coding aided D2D communications. *IEEE Trans Mob Comput* 16:3459–3472

Relays in LTE-Advanced

- [Relaying in LTE-Advanced](#)

Reliable

- [Smart Grid Communication Based on IEEE 2030 Standard](#)

Remote Health Sensing Using WiFi Signal

- [Sleep Monitoring Using WiFi Signal](#)

Repulsive Point Process

- [Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes](#)

Resolution Mapping

- [Universal Identifier-Based Network](#)

Resource Allocation

- [Bandwidth Allocation in Data Center Networks](#)
- [Machine Learning Paradigms in Wireless Network Association](#)
- [Resource Allocation in UAV-Aided Wireless Networks](#)

Resource Allocation in Industrial IoT

Ling Lyu

School of Information Science and Technology,
Dalian Maritime University, Dalian, China

Synonyms

[Industrial cyber-physical systems](#); [Resource assignment](#); [Resource management](#); [Resource scheduling](#)

Definition

Resource allocation in Industrial IoT is the joint design and optimization of network resources in multiple domains (frequency, time, power, and so on) to provide better services for different industrial applications.

Historical Background

The Internet of Things (IoT) provides a bridge between sensors, controllers, and actuators to enable the ultra-reliable low-latency communication services for control applications in industrial automation (Zhu et al. [2014](#); Lyu et al. [2016](#), [2018a](#); Stojmenovic [2014](#); Cheng et al. [2016](#), n.d.; Chen et al. [2015](#)). The emerging IoT has a wide range of applications in many fields, such as manufacturing, oil and gas, energy mining, metals, and other industrial sectors (Lyu et al.

2017; Zhang et al. 2017a, b; Guo et al. 2017; Lyu et al. 2018b).

Industrial IoT can improve interconnectivity, flexibility, scalability, time efficiency, cost effectiveness, security, productivity, and operational efficiency in the industries (Raza et al. 2017). IoT can also serve as a platform to establish intelligent network of devices (Gubbi et al. 2013) which can interrelate data and processes to effectively establish feedback control systems for industrial automation.

According to Li et al. (2018), IoT will support the integration of approximately 5.5 million things every day by 2016, with estimation of 20.8 billion things by 2020. The rapid increase of IoT machines brings some technical challenges in order to meet the complex requirements. For example, the existing wireless communication technologies with low power are still under investigation to meet the energy efficient requirement with high system reliability. Therefore, research efforts should be carried out to tackle the obstacle of industrial IoT to be deployed in practice, especially in industrial applications (Al-Rubaye et al. 2017). Industrial IoT is a specific branch of IoT which addresses the communication and connectivity issues in industrial environments and offers suitable solutions. Among all possible wireless communication network solutions for IoT, 5G is considered as a promising technology, which integrates entities, communications, and control technologies (Atat et al. 2017).

Since 5G supports enhanced mobile broadband communications, ultra-reliable and low-latency communications (URLLC), and massive machine-type communications (mMTC), it is clear that URLLC and mMTC in 5G are closely related to industrial IoT. Thus, 5G provides an ideal platform for communications in industrial IoT. In the application area of industrial automation, industrial IoT performs system monitoring and control with the assistance of 5G systems (Holfeld et al. 2016; Shariatmadari et al. 2015). In such a 5G-enabled industrial IoT, one major technical challenge is the resource allocation in industrial IoT, which takes charge of designing resource-efficient communication to facilitate various resources while providing

the satisfactory quality of service as well as supporting dynamic network environments and control systems.

Foundations

For the industrial IoT, the control system stability is severely affected by the wireless transmission reliability, which in turn relies heavily on the scheduling of limited wireless communication resources, such as channel, time, and power. The control system stabilization is regarded as the quality of performance requirement in the application layer, which can be mapped into a certain quality of service requirement in the lower layers. In this way, the desired quality of performance requirement can be achieved with proper resource allocation schemes, i.e., wireless transmission is regarded as another degree of freedom to be optimized in the industrial IoT system. Therefore, it is crucial to investigate the resource allocation for industrial IoT to improve the control-communication performance.

Therefore, it is important to achieve a dynamic and efficient resource allocation to cope with packet loss, delay, and other serious issues. To address this issue, the following foundations are necessary.

Dynamic Priority-Based Resource Allocation

In general, an industrial process includes multi-class traffics, such as emergency traffic, regulatory control traffic, alerting traffic, monitoring traffic, and so on. Different traffic usually has different importance. In this regard, the priority-based communication provides a better chance at the effective data transmission. However, the general use of predefined traffic priority may not be a suitable solution, since the priority of traffic in industrial networks changes on runtime, depending on variations in sensor readings. Hence, the widely proposed priority-based MAC protocols are inefficient in the scenario where certain variation in the existing priority may occur. As a solution, a more effective way to deal with processes like these is by establishing dynamic

priority, which could prioritize in real time. The concept behind dynamic priority scheduling is to form adaptive and evidence-based decision of which information is more critical and must be delivered on priority. Inclusion of dynamic priority scheduling meets the challenges of time constraint and information priority scheduling by efficiently sacrificing the resource of low priority data traffics.

Energy Efficiency

As many industrial IoT applications need to run for years on batteries, it creates a demand for energy-efficient designs. To complement such designs, upper-layer approaches play important roles through energy-efficient operation. Many energy-efficient schemes have been proposed in recent years (Rault et al. 2014), but those approaches are not immediately applicable to industrial IoT. Industrial IoT applications typically need a dense deployment of numerous devices. Sensed data can be sent in queried form or in a continuous form which in a dense deployment can consume a significant amount of energy. Green networking is thus crucial in industrial IoT to reduce power consumption and operational costs. It will lessen pollution and emissions and make the most of surveillance and environmental conservation. Although there are numerous methods to achieve energy efficiency, such as using lightweight communication protocols or adopting low-power radio transceivers as described above, the recent technology trend in energy harvesting provides another fundamental method to prolong battery life. Thus, energy harvesting is a promising approach for the emerging industrial IoT.

Timing and Reliability

Industrial IoT devices are typically deployed in noisy environments for supporting mission- and safety-critical applications, and have stringent timing and reliability requirements on timely collection of environmental data and proper delivery of control decisions. The QoS offered by IIoT is thus often measured by how well it satisfies the end-to-end deadlines of the real-time sensing and control tasks executed in the system (Ferrari et al.

2018; Zheng et al. 2016). Time-slotted packet scheduling in industrial IoT plays a critical role in achieving the desired QoS. For example, many industrial IoT systems perform network resource allocation via static data link layer scheduling (Leng et al. 2014; Saifullah et al. 2015) to achieve deterministic end-to-end real-time communication. Such approaches typically take a periodic approach to gather the network health status, and then distribute the updated network schedule information. This process however is slow, not scalable, and incurs considerable network overhead. The explosive growth of industrial IoT applications especially in terms of their scale and complexity has dramatically increased the level of difficulty in ensuring the desired real-time performance. The fact that most industrial IoT must deal with unexpected disturbances further aggravate the problem.

Key Applications

In general, industrial network system consists of spatially distributed sensors and multiple edge and/or remote estimators/controllers. Sensors cooperatively monitor physical system states or environmental conditions and estimators perform situation awareness and system control in a coordinated way.

Situation Awareness

One main application of resource allocation in the industrial IoT is the situation awareness, which aims to estimate the state parameters to understand the situation of sensed environment. However, sensors are generally powered with batteries and wireless spectrum resource is scarce. The limited energy and spectrum resource makes the functionalities of industrial IoT systems on sensing, transmitting, and control quite restricted. The potential contradictions between estimation accuracy and resource limitation motivate the investigation of resource allocation and state estimation co-design strategy. Therefore, it is important and necessary to properly assign various resources under the consideration of state estimation performance to improve the estimation performance in

a resource-efficient way, which could be achieved by investigating an energy-efficient communication and estimation co-design strategy for industrial IoT.

However, for the considered monitoring application in IIoT systems, the measurements obtained by nearby sensors have much similar information, resulting in high measurement-correlation. In addition, a large number of sensors, which are deployed to achieve the ubiquitous situation awareness, may cause problems in the connection establishment and radio resource allocation. In order to provide the required quality of experience for the monitoring application, besides the usually considered energy consumption, more application-centric performance metrics should be taken into account. However, this is challenging in the dynamic industrial environment, since wireless signals may suffer from severe fading, shadowing, multi-path propagation, and interference, due to obstructions from big metal equipments, movements of machineries, electromagnetic interferences, high temperatures and humidities, and so on (Vitturi et al. 2013).

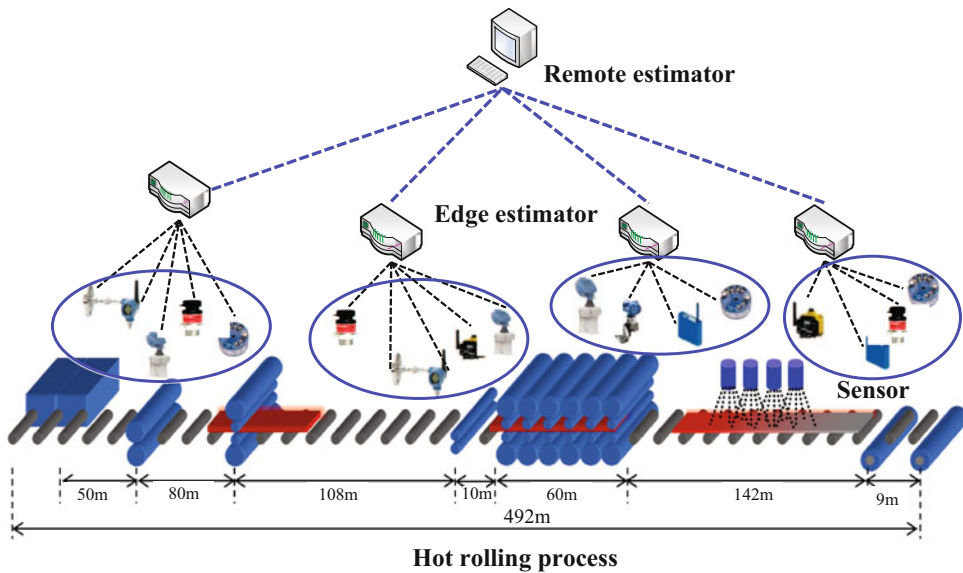
Recently, the hierarchical framework combining with centralized and distributed state esti-

mations has been studied for guaranteeing the prescribed estimation accuracy with limited communication resources. A hierarchical fusion estimator presented in Song et al. (2014) involves an optimal local estimator in the minimum variance sense and a fusion strategy combining local estimates with the previous fused estimate. In order to provide a satisfactory estimation precision with low computational burden, Zhang et al. (2016) present a sequential measurement fusion strategy for local estimators and a sequential covariance intersection fusion strategy for fusion estimation.

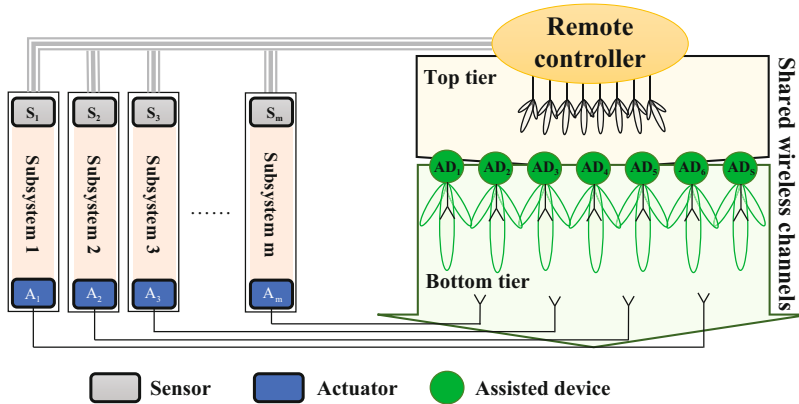
For the industrial IoT, many MTC devices are deployed to achieve the ubiquitous situation awareness in a coordinated way. The similar fog-cloud hierarchical architecture consisting of many sensors (MTC devices), several edge estimators (MTC APs), and one remote estimator (MTC gateway) shown in Fig. 1, where sensors are divided into clusters and each cluster includes a cluster head acting as the edge estimator.

System Control

In industrial IoT, the separation of control and actuation involves the data transmission of control commands over shared wireless



Resource Allocation in Industrial IoT, Fig. 1 Hierarchical network architecture for state estimation



Resource Allocation in Industrial IoT, Fig. 2 Network architecture for system control in industrial IoT

channels. In general, the more information the controller sends, the more precise actuation becomes, leading to that the control performance improves at the cost of more energy and spectrum resources. In the past decade, the communication-constrained control has been widely investigated in the area of networked control (Donkers et al. 2011; Cao et al. 2015; Gatsis et al. 2015; Al-Areqi et al. 2015). In order to guarantee the system stability and improve control performance, the control-scheduling strategy is adjusted based on system dynamics, while satisfying the resource constraint imposed by the communication network.

Besides designing the control-scheduling strategy based on system dynamics, another important issue is how to successfully complete the resource allocation for the scheduled subsystems over lossy wireless channels. However, in these aforementioned studies, the assumption about the perfect communication of control commands is over ideal. In practice, the harsh industrial wireless environments, such as electromagnetic interference, high temperature and humidity levels, vibrations, dirt, and dust, may result in the uncertain and stochastic wireless channels (Gungor and Hancke 2009; Lei et al. 2016; Doe 2004). How to achieve a reliable and efficient data transmission with poor channel conditions has attracted a lot of attention from both academia and industry. Traditional techniques to enhance the reliability of wireless

communications, such as retransmission and redundancy (Mahmood et al. 2015), focus on transmitting additional packets or bits to achieve the loss recovery. A price we need to pay, however, is the increasing time delay and energy consumption. Thus, they are not suitable for the delay-sensitive and energy-limited industrial IoT applications (Sadi and Ergen 2015). When the time diversity is not viable in the delay-sensitive scenario, exploiting the spatial diversity is a real choice to improve the transmission reliability for industrial automation systems (Swamy et al. 2017). Swamy et al. (2017) has proposed a wireless communication protocol that capitalizes on multiuser diversity and cooperative communication to achieve the ultra-reliability with a low-latency constraint.

In general, multiple independent subsystems share wireless channels. Each control subsystem consists of one sensor-actuator pair, whose feedback control commands are transmitted from the remote controller to actuators. In addition, the remote controller transmits many independent control signals to many independent actuators. As shown in Fig. 2, the dynamics of different subsystems are generally different, which makes the control inputs of different subsystems different. In this case, different actuators usually execute different control commands to stabilize their own subsystems. For control applications in industrial IoT with multiple subsystems, subsystems with different dynamic characteristics have distinct

impacts on guaranteeing the stability and reducing the overall control cost of whole system. To solve such problem, it is necessary to design an efficient resource allocation strategy to guarantee the system stability and reduce the overall cost. By taking the unreliable transmission imposed by the lossy wireless channels into account, the remote controller is in charge of jointly optimizing the control scheduling and the resource allocation, which is beneficial to guarantee the stability of each subsystem constrained by limited spectrum and energy resources.

Conclusion

The resource allocation strategy severely affects the data transmission reliability and the control system performance. Therefore, it is necessary to properly assign multidomain resources for industrial IoT to improve the control-communication performance.

Cross-References

- [Mobile Edge Computing: Low Latency and High Reliability](#)

References

- Al-Areqi S, Gorges D, Liu S (2015) Event-based networked control and scheduling codesign with guaranteed performance. *Automatica* 57:128–134
- Al-Rubaye S, Kadhum E, Ni Q, Anpalagan A (2017) Industrial internet of things driven by sdn platform for smart grid resiliency. *IEEE Internet Things J* PP(99): 1–11
- Atat R, Liu L, Chen H, Wu J, Li H, Yi Y (2017) Enabling cyberphysical communication in 5g cellular networks: challenges, spatial spectrum sensing, and cyber-security. *IET Cyber-Physical Syst Theory Appl* 2(1):49–54
- Cao X, Zhou X, Liu L, Cheng Y (2015) Energy-efficient spectrum sensing for cognitive radio enabled remote state estimation over wireless channels. *IEEE Trans Wirel Commun* 14(4):2058–2071
- Chen C, Yan J, Lu N, Wang Y, Yang X, Guan X (2015) Ubiquitous monitoring for industrial cyber-physical system over relay assisted wireless sensor networks. *IEEE Trans Emerg Top Comput* 3(3):352–362
- Cheng N, Lu N, Zhang N, Yang T, Shen X, Mark JW (2016) Vehicle-assisted device-to-device data delivery for smart grid. *IEEE Trans Veh Technol* 65(4): 2325–2340
- Cheng N, Lyu F, Chen J, Xu W, Zhou H, Zhang S, Shen X (2016) Big data driven vehicular networks. *IEEE Netw*. <https://arxiv.org/pdf/1804.04203.pdf>
- Doe U (2004) Industrial wireless technology for the 21st century. Report, Technology Foresight, Winter
- Donkers MCF, Heemels WPMH, Wouw NVD, Hetel L (2011) Stability analysis of networked control systems using a switched linear systems approach. *IEEE Trans Autom Control* 56(9):2101–2115
- Ferrari P, Flammini A, Sisinni E, Rinaldi S, Brandão D, Rocha MS (2018) Delay estimation of industrial IoT applications based on messaging protocols. *IEEE Trans Instrum Meas* 67(9):1–12
- Gatsis K, Pajic M, Ribeiro A, Pappas GJ (2015) Opportunistic control over shared wireless channels. *IEEE Trans Autom Control* 60(12):3140–3155
- Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (iot): a vision, architectural elements, and future directions. *Futur Gener Comput Syst* 29(7):1645–1660
- Gungor VC, Hancke GP (2009) Industrial wireless sensor networks: challenges, design principles, and technical approaches. *IEEE Trans Ind Electron* 56(10): 4258–4265
- Guo H, Ren J, Zhang D, Zhang Y, Hu J (2017) A scalable and manageable IoT architecture based on transparent computing. *J Parallel Distrib Comput*. <https://doi.org/10.1016/j.jpdc.2017.07.003>
- Holfeld B, Wieruch D, Wirth T, Thiele L, Ashraf SA, Huschke J, Aktas I, Ansari J (2016) Wireless communication for factory automation: an opportunity for LTE and 5G systems. *IEEE Commun Mag* 54(6):36–43
- Lei L, Kuang Y, Shen X, Yang K, Qiao J, Zhong Z (2016) Optimal reliability in energy harvesting industrial wireless sensor networks. *IEEE Trans Wirel Commun* 15(8):5399–5413
- Leng Q, Wei Y-H, Han S, Mok AK, Zhang W, Tomizuka M (2014) Improving control performance by minimizing jitter in rt-wifi networks. In: 2014 IEEE Real-Time Systems Symposium (RTSS). IEEE, pp 63–73
- Li S, Ni Q, Sun Y, Min G, Al-Rubaye S (2018) Energy-efficient resource allocation for industrial cyber-physical iot systems in 5g era. *IEEE Trans Ind Inf* 14(6):2618–2628
- Lyu L, Chen C, Yan J, Lin F, Hua C, Guan X (2016) State estimation oriented wireless transmission for ubiquitous monitoring in industrial cyber-physical systems. *IEEE Trans Emerg Top Comput*. <https://doi.org/10.1109/TETC.2016.2573719>
- Lyu L, Chen C, Hua C, Zhu S, Guan X (2017) Co-design of stabilisation and transmission scheduling for wireless control systems. *IET Control Theory Appl* 11(11):1767–1778

- Lyu L, Chen C, Zhu S, Guan X (2018a) 5G enabled code-sign of energy-efficient transmission and estimation for industrial IoT systems. *IEEE Trans Ind Inf* 14(6): 2690–2704
- Lyu L, Chen C, Zhu S, Cheng N, Tang Y, Guan X, Shen XS (2018b) Dynamics-aware and beamforming-assisted transmission for wireless control scheduling. *IEEE Trans Wirel Commun*. <https://doi.org/10.1109/TWC.2018.2869589>
- Mahmood MA, Seah WK, Welch I (2015) Reliability in wireless sensor networks: a survey and challenges ahead. *Comput Netw* 79:166–187
- Rault T, Bouabdallah A, Challal Y (2014) Energy efficiency in wireless sensor networks: a top-down survey. *Comput Netw* 67:104–122
- Raza M, Aslam N, Le-Minh H, Hussain S, Cao Y, Khan NM (2017) A critical analysis of research potential, challenges, and future directives in industrial wireless sensor networks. *IEEE Commun Surv Tutor* 20(1):39–95
- Sadi Y, Ergen SC (2015) Energy and delay constrained maximum adaptive schedule for wireless networked control systems. *IEEE Trans Wirel Commun* 14(7):3738–3751
- Saifullah A, Xu Y, Lu C, Chen Y (2015) End-to-end communication delay analysis in industrial wireless networks. *IEEE Trans Comput* 64(5):1361–1374
- Shariatmadari H, Ratasuk R, Iraj S, Laya A, Taleb T, Jäntti R, Ghosh A (2015) Machine-type communications: current status and future perspectives toward 5g systems. *IEEE Commun Mag* 53(9):10–17
- Song H, Zhang W, Yu L (2014) Hierarchical fusion in clustered sensor networks with asynchronous local estimates. *IEEE Signal Process Lett* 21(12):1506–1510
- Stojmenovic I (2014) Machine-to-machine communications with in-network data aggregation, processing and actuation for large scale cyber-physical systems. *IEEE Internet Things J* 1(2):122–128
- Swamy VN, Suri S, Rigge P, Weiner M, Ranade G, Sahai A, Nikolic B (2017) Real-time cooperative communication for automation over wireless. *IEEE Trans Wirel Commun* 16(11):7168–7183
- Vitturi S, Tramarin F, Seno L (2013) Industrial wireless networks: the significance of timeliness in communication systems. *IEEE Ind Electron Mag* 7(2):40–51
- Zhang W, Chen B, Chen MZQ (2016) Hierarchical fusion estimation for clustered asynchronous sensor networks. *IEEE Trans Autom Control* 61(10):3064–3069
- Zhang H, Liu X, Li C, Chen Y, Che X, Wang LY, Lin F, Yin G (2017a) Scheduling with predictable link reliability for wireless networked control. *IEEE Trans Wirel Commun* 16(9):6135–6150
- Zhang N, Yang P, Zhang S, Chen D, Zhuang W, Liang B, Shen X (2017b) Software defined networking enabled wireless network virtualization: challenges and solutions. *IEEE Netw* 31(5):42–49
- Zheng T, Gidlund M, Åkerberg J (2016) Wirarb: a new mac protocol for time critical industrial wireless sensor network applications. *IEEE Sensors J* 16(7):2127–2139

- Zhu S, Chen C, Ma X, Yang B, Guan X (2014) Consensus based estimation over relay assisted sensor networks for situation monitoring. *IEEE J Sel Top Sign Proces* 9(2):1–14

Resource Allocation in NOMA-Assisted D2D Networks

Yanpeng Dai

Xidian University, Xi'an, China

Synonyms

Device-to-device (D2D) communication; Interference management; Non-orthogonal multiple access (NOMA); Resource management

Definition

Resource allocation in NOMA-assisted D2D networks is to assign the available resources including D2D communication modes, spectrum resources, and transmit power.

Historical Background

The fifth-generation (5G) network will possess an architecture of heterogeneous networks integrating with multiple advanced communication technologies to satisfy high performance requirements (Agiwal et al. 2016). Non-orthogonal multiple access (NOMA) technology is widely recognized as a key technology for 5G networks to support massive connectivity and high network capacity (Ding et al. 2017). The NOMA technology can utilize the non-orthogonal domain to enable that multiple users reuse the same orthogonal resource to transmit their signals, which effectively improves the spectrum utilization. The current NOMA technologies mainly include two

categories, which are based on the power domain and the code domain, respectively. Specifically, for the power domain NOMA, the signals of different users are transmitted on different power levels. Then, the successive interference cancellation (SIC) is utilized to decode the signals of the users by the difference on power domain (Islam et al. 2017). For code domain NOMA, several techniques have been proposed by the academia and industry, including sparse code multiple access (SCMA), pattern division multiple access, low-density spreading, and so on. By the code domain NOMA, the user is assigned with unique codeword or pattern to transmit the signal. At receiver, the multi-user joint detection is utilized to decode the signals of users according to the design of codebooks or pattern clusters, e.g., message passing algorithm for SCMA (Nikopour and Baligh 2013; Taherzadeh et al. 2014; Dai et al. 2017).

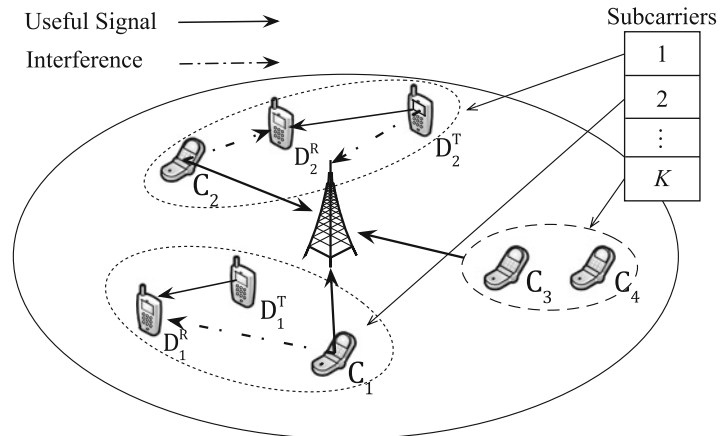
Device-to-device communication is one of the promising technologies to provide the proximity service for wireless devices, e.g., smartphones, connected vehicles, and drones (Asadi et al. 2014). Exploiting D2D communication, the nearby users can directly communicate with each other without the relay of base station (BS), which effectively improves the network capacity and energy efficiency. Different from other short-range communication techniques, D2D communications coexist with cellular networks and are managed by the cellular network. There are two D2D communication modes, which are the over-

lay mode and the underlay mode (Yu et al. 2014). In the overlay mode, D2D users dedicatedly utilize the network resource which is orthogonal with the cellular resources. Hence, beside the cellular resources, the dedicated resources should be required for D2D communication. In the underlay mode, D2D users share the same resource with cellular users to promote the resource utilization. However, the mutual interference between cellular users and D2D users will be caused, which degrades the performance of both cellular transmission and D2D transmission (Feng et al. 2013). Therefore, it is necessary to implement the efficient resource allocation for D2D communications and cellular users. Furthermore, it is significant to design the resource allocation and network scheduling for D2D in NOMA-enhanced cellular network, which is an important scenario in 5G networks. Figure 1 shows an example of NOMA-assisted cellular and D2D networks.

Since multiple cellular users transmit the signals via the same resource in NOMA system, the co-channel interference exists between the cellular users, which is different from traditional orthogonal multiple access (OMA) system (Yang et al. 2016; Ding et al. 2014; Di et al. 2016). Therefore, we should consider the impact of D2D communication on multiple users' NOMA transmission. For example, in power domain NOMA, the underlay D2D transmission can cause the interference to NOMA transmissions, which changes the co-channel interference and further influences the original execution of SIC

R

Resource Allocation in NOMA-Assisted D2D Networks, Fig. 1 D2D in NOMA uplink cellular network



decoding. The failure of SIC decoding leads to the series of failed user detection rather than only one user to degrade the reliability of NOMA system. Therefore, in NOMA-enhanced cellular networks, the mode selection and radio resource allocation for D2D communication is an important and challenging issue, which can acquire the advantages of NOMA and D2D on the spectrum efficiency and network connectivity (Pan et al. 2018; Liu et al. 2016).

Foundations

For NOMA-assisted D2D networks, the resource allocation mainly focuses on the mode selection of D2D transmission and resource management. The mode selection is to determine the resource sharing approach of D2D transmission in NOMA cellular users. Due to the non-orthogonal resource sharing between cellular users, the mode selection should consider the impact of the D2D transmission on NOMA transmissions to guarantee the efficiency of resource sharing between D2D transmission and NOMA cellular users. The radio resource management includes the subchannel/codebook assignment and power allocation, which aims to improve the spectrum efficiency of NOMA-assisted D2D networks.

Mode Selection for NOMA-Assisted D2D Communications

There are two typical D2D communications, the overlay mode and the underlay mode. The mode selection between the overlay mode and underlay mode has been widely studied by a number of literature (Lin et al. 2014; ElSawy et al. 2014; Ye and Zhang 2015). However, the D2D mode selection in NOMA system is a different issue and becomes more challenging. In power domain NOMA, if we select the underlay mode for the D2D transmission, the D2D transmission cannot break the NOMA transmissions of multiple cellular users on the same resource, which is different from that in OMA system where the D2D transmission only interfere with only one cellular user (Pan et al. 2018; Zhao et al. 2017). For SCMA-assisted D2D and cellular hybrid network, the

D2D mode selection depends on the mapping rules between the users and SCMA codebooks, which become more complicated than traditional OMA-based hybrid network (Zhai et al. 2017). In fact, the overlay mode and underlay mode are proposed based on the features of OMA system. Due to exploiting the code domain and power domain, the NOMA system has the potential to provide a new D2D mode to improve the efficiency of resource sharing between the cellular user and D2D pair. The NOMA-based D2D communication is promising and challenging issues for the future research works.

Resource Management for D2D Communications in NOMA System

For D2D communications, the radio resource management is of critical importance to guarantee the quality of service (QoS) of the network and satisfy the user requirements. Efficient resource management can restrain the interference, reduce energy consumption, and increase system rate. Due to the coexisting of D2D users and cellular users, the interference management is a challenging issue, which becomes more complicated in NOMA-enhanced D2D and cellular hybrid networks. In NOMA-enhanced D2D and cellular hybrid network, there are two types of the interference, which are the co-channel interference between NOMA users and the cross-tier interference between D2D users and cellular users. Therefore, how to design efficient resource allocation scheme is an important issue for interference management in this hybrid networks. The joint optimization of subchannel assignment and power allocation is a vital resource management method to alleviate the interference and improve the system capacity. For power domain NOMA, the joint subchannel and power allocation has been investigated by several works (Pan et al. 2018; Zhao et al. 2017). For code domain NOMA, the resource allocation is also studied by several papers (Zhai et al. 2017; Dai et al. 2016). The resource management issue for NOMA-assisted D2D networks is mainly formulated and solved by the optimization theory (Boyd and Vandenberghe 2004), matching game theory (Gu et al. 2015), and graph theory (West 2001).

Key Applications

Recently, the NOMA has been investigated in many 5G scenario to break through the bottleneck of the traditional OMA system, especially for ultra dense network and the scenario demanding the massive connectivity. Similarly, D2D communications are applied into emerging scenarios, such as the wireless caching, cooperative transmission, and connected vehicular networks (Qian et al. 2017; Di et al. 2017; Ding et al. 2018). Therefore, the resource allocation for D2D communications in NOMA-enhanced wireless communication system plays a vital rule for improving network capacity and user experience. We give several applications of the resource allocation in NOMA-assisted D2D networks as follows.

- **QoS-aware network capacity maximization:** One main application of resource allocation in NOMA-assisted D2D network is to maximize the network capacity while satisfying the QoS requirement of the users (Pan et al. 2018; Zhao et al. 2017). The QoS requirements have several aspects including the minimum rate, maximum transmission delay, and maximum energy consumption. However, the complicated interference between transmission links can not only severely degrade the network capacity but also destroy the original NOMA cellular transmissions. Thus, the efficient resource allocation should coordinate the interference to improve the network capacity while meeting the user requirements. The resource allocation problem usually can be formulated as an optimization problem.
- **Transmission delay minimization:** Due to the advantages on transmission efficiency, the NOMA technology is utilized to reduce the transmission delay. In the Internet of Things scenarios, the D2D communication enables the information transmission between the devices, such as the vehicle-to-vehicle communications and e-health sensing, where the transmission delay and link reliability are focused on. Thus, we should design the

appropriate resource allocation scheme for this kind of NOMA-enabled D2D networks to achieve low-latency and high-reliable communication. The NOMA for high-reliable and low-latency vehicular communications has been studied in Di et al. (2017). The performance of NOMA-enabled wireless caching is investigated in Ding et al. (2018).

Cross-References

- [Resource Management for Heterogeneous Networks](#)
- [Non-orthogonal Multiple Access](#)
- [Resource Allocation in NOMA-Assisted D2D Networks](#)

References

- Agiwal M, Roy A, Saxena N (2016) Next generation 5G wireless networks: a comprehensive survey. *IEEE Commun Surv Tutor* 18(3):1617–1655
- Asadi A, Wang Q, Mancuso V (2014) A survey on device-to-device communication in cellular networks. *IEEE Commun Surv Tutor* 16(4):1801–1819
- Boyd S, Vandenberghe L (2004) *Convex optimization*. Cambridge University Press, Cambridge
- Dai Y, Sheng M, Zhao K, Liu L, Liu J, Li J (2016) Interference-aware resource allocation for D2D underlaid cellular network using SCMA: a hypergraph approach. In: 2016 IEEE WCNC, Apr 2016, pp 1–6
- Dai Y, Sheng M, Liu J, Liu L, Li J (2017) Backtracking codebook matching algorithm based on codebook correlation for uplink SCMA system. *IEEE Commun Lett* 21(7):1597–1600
- Di B, Song L, Li Y (2016) Radio resource allocation for uplink sparse code multiple access (SCMA) networks using matching game. In: *Proceedings of IEEE ICC*, May 2016, pp 1–6
- Di B, Song L, Li Y, Li GY (2017) Non-orthogonal multiple access for high-reliable and low-latency V2X communications in 5G systems. *IEEE J Sel Areas Commun* 35(10):2383–2397
- Ding Z, Yang Z, Fan P, Poor HV (2014) On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Signal Process Lett* 21(12):1501–1505
- Ding Z, Liu Y, Choi J, Sun Q, El-kashlan M, Chih-Lin I, Poor HV (2017) Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun Mag* 55(2):185–191
- Ding Z, Fan P, Karagiannis GK, Schober R, Poor HV (2018) NOMA assisted wireless caching: strate-

- gies and performance analysis. *IEEE Trans Commun* 66(10):4854–4876
- ElSawy H, Hossain E, Alouini M (2014) Analytical modeling of mode selection and power control for underlay d2d communication in cellular networks. *IEEE Trans Commun* 62(11):4147–4161
- Feng D, Lu L, Wu YY, Li GY, Feng G, Li S (2013) Device-to-device communications underlying cellular networks. *IEEE Trans Commun* 61(8):3541–3551
- Gu Y, Saad W, Bennis M, Debbah M, Han Z (2015) Matching theory for future wireless networks: fundamentals and applications. *IEEE Commun Mag* 53(5):52–59
- Islam SMR, Avazov N, Dobre OA, Kwak KS (2017) Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. *IEEE Commun Surv Tutor* 19(2):721–742
- Lin X, Andrews JG, Ghosh A (2014) Spectrum sharing for device-to-device communication in cellular networks. *IEEE Trans Wirel Commun* 13(12):6727–6740
- Liu J, Sheng M, Liu L, Shi Y, Li J (2016) Modeling and analysis of SCMA enhanced D2D and cellular hybrid network. *IEEE Trans Commun* 65(1):173–185
- Nikopour H, Baligh H (2013) Sparse code multiple access. In: *Proceedings of IEEE PIMRC*, London, Sept 2013, pp 332–336
- Pan Y, Pan C, Yang Z, Chen M (2018) Resource allocation for D2D communications underlying a NOMA-based cellular network. *IEEE Wirel Commun Lett* 7(1):130–133
- Qian L, Wu Y, Zhou H, Shen X (2017) Dynamic cell association for non-orthogonal multiple-access V2S networks. *IEEE J Sel Areas Commun* 35(10):2342–2356
- Taherzadeh M, Nikopour H, Bayesteh A, Baligh H (2014) Scma codebook design. In: *Proceedings of IEEE VTC Fall*, Vancouver, Sept 2014, pp 1–5
- West DB (2001) *Introduction to graph theory*, vol 2. Prentice hall Upper Saddle River, New Jersey
- Yang Z, Ding Z, Fan P, Al-Dhahir N (2016) A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems. *IEEE Trans Wirel Commun* 15(11):7244–7257
- Ye J, Zhang YJ (2015) A guard zone based scalable mode selection scheme in D2D underlaid cellular networks. In: *2015 IEEE ICC*, June 2015, pp 2110–2116
- Yu G, Xu L, Feng D, Yin R, Li GY, Jiang Y (2014) Joint mode selection and resource allocation for device-to-device communications. *IEEE Trans Commun* 62(11):3814–3824
- Zhai D, Sheng M, Wang X, Sun Z, Xu C, Li J (2017) Energy-saving resource management for D2D and cellular coexisting networks enhanced by hybrid multiple access technologies. *IEEE Trans Wirel Commun* 16(4):2678–2692
- Zhao J, Liu Y, Chai KK, Chen Y, El Kashlan M (2017) Joint subchannel and power allocation for NOMA enhanced D2D communications. *IEEE Trans Commun* 65(11):5081–5094

Resource Allocation in SDN/NFV-Enabled 5G Networks

Junling Li¹, Weisen Shi¹, and Ning Zhang^{2,3,4}

¹University of Waterloo, Waterloo, ON, Canada

²Peng Cheng Laboratory, Shenzhen, China

³Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

⁴College of Intelligence and Computing, Tianjin University, Tianjin, China

Synonyms

Resource management; Resource planning; Resource slicing

Definitions

Resource allocation is the assignment of available resources including computing, storage, and networking resources to the customized services to achieve better QoS provisioning. In SDN/NFV-enabled 5G networks, heterogeneous resources are virtualized and allocated to different network services composed by ordered chains of virtual network functions (VNFs).

Historical Background

Traditionally, network infrastructure are managed and maintained by different network operators independently. In this business model, to deploy a network service for each network operator, the data traffic needs to flow through a certain fixed group of middleboxes, such as firewall, DNS (Domain Name System), and IDS (Intrusion Detection System), in a predetermined order (Li et al. 2018). The problem of choosing the set of required middleboxes and deciding the corresponding routing paths is called “middleboxes orchestration.” In an existing network, this task is manually executed through a “trial-and-error” approach, which is very costly and time-consuming. Furthermore, once placed, it is costly and impractical to change the location of a

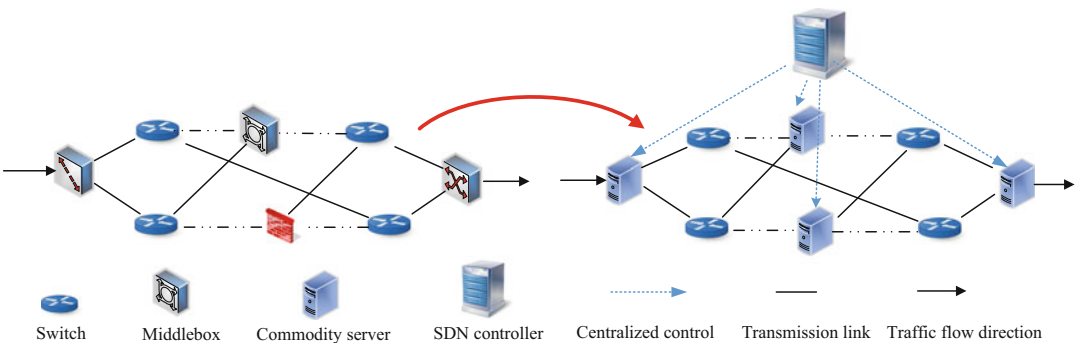
middlebox according to the variations of network conditions.

In the context of SDN/NFV-enabled 5G networks, a network service can be formed with a set of chained virtual network functions (VNFs). Therefore, a network service is typically defined by the types of the VNFs chained, the processing order of the VNFs in the chain, and the embedding of the chain onto the network function virtualization infrastructure (NFVI). Figure 1 illustrates the transition from traditional network to SDN/NFV-enabled 5G network. Note that the function-specific middleboxes are replaced by commodity servers to host VNFs. A logically centralized SDN controller, which has a global view of the whole network, is used to manage and allocate the resources in the network. Through virtualizing network functions, new services can be accommodated by updating the software on the network servers instead of deploying new devices. In this way, the service provisioning cost can be reduced while the resource utilization can be improved. On the other hand, the use of SDN facilitates the realization of network openness and programmability. Despite the benefits brought by NFV and SDN, one of the main challenges in SDN/NFV-enabled 5G networks is, how to efficiently manage, allocate, and optimize the heterogeneous resources, including computing, storage, and networking resources, to the customized services for better QoS provisioning (Li et al. 2017).

Foundations

In SDN/NFV-enabled 5G networks, the resource allocation task (also referred to as network function virtualization resource allocation, NFV-RA) is implemented in three aspects (Herrera and Botero 2016), i.e., given the traffic metrics, QoS requirements, and resource constraints of the substrate network, we first need to compose a VNF chain to form a virtual network (VN) topology. Then virtual resources allocated to each element (VNF or virtual link) need to be optimized in order to minimize the provisioning cost and to satisfy the QoS requirements. This process is usually referred to as *VNF chain composition*. Next, we need to embed the virtual network (i.e., the VNF chain) onto the substrate network. During this stage, we need to assign the physical resources in an economical way to meet the resource demands of virtual nodes and links. This involves determining the locations of NFV nodes to host the VNFs and the routing from source to the destination. This process is referred to as *VNF chain embedding*. Finally, we need to schedule the VNFs for multiple services to minimize the service completion time and maximize the network performance. In other words, when multiple VNFs are embedded on the same NFV node, each of them should be assigned an optimal time slot. This process is referred to as *VNF scheduling* (Xie et al. 2016).

R



Resource Allocation in SDN/NFV-Enabled 5G Networks, Fig. 1 Global resource management and allocation via centralized SDN controller in core networks

- VNF chain composition: In the NFV environment, a network service request is an entity composed by a number of VNFs in a pre-determined order (Martini et al. 2015). Different from the existing network where network functions are run on dedicated hardware appliances on fixed locations, VNFs are softwares with much higher deployment flexibility, leading to increased complexity in composing the VNF chains. Therefore, it is crucial to investigate the chain composition process to produce VNF chains more efficiently and to allow the fast deployment of customized SDN/NFV-enabled network services (Ye et al. 2016).
- For a given network service request, the following information is typically known: (1) set of VNFs in the request and the dependencies between them; (2) fixed substrate nodes of initial and terminal points for the traffic flow; (3) initial data rate of the traffic flow; (4) the ratio of outgoing to incoming data rate of each VNF; and (5) the processing capacity demand per bandwidth unit of each VNF. There are many factors that increase the complexity and hardness of the VNF chain composition problem. First of all, the ordering of VNFs are flexible but has to respect to the dependencies between the VNFs. Different ordering of VNFs may result in different physical resources requirements (which is related to the embedding cost) for the resultant VNF chain. This flexibility increases the searching space of the optimal VNF chaining. Secondly, traffic splitting may be required in situations where the residual physical resources of a node (or a link) are not enough to support the network service. In such scenarios, some VNF has to be instantiated on multiple NFV-enabled nodes and the data traffic has to traverse through different VNF instances running on multiple nodes. This will further increase the hardness of the VNF chain composition problem. Thirdly, once the order of VNFs is determined, we need to determine the set of physical resources allocated to the VNFs and virtual links, to minimize the embedding cost while providing QoS-guaranteed services. Lastly, as the physical resources of the substrate network is limited, it is also important to take the substrate network condition, more specifically, the locations, residual capacities, and host abilities of substrate nodes into consideration in the VNF chaining process to improve the overall network performance. This will make it even harder to obtain the optimal solution to the VNF chain composition problem.
- VNF chain embedding: Once the VNF chain is composed via the chain composition process, the input to the second stage of NFV-RA is obtained. The VNF chain embedding process tries to place the VNFs on NFV nodes in the NFVI with multiple network service requests consideration. The VNFs should be consolidated in a way such that the physical resources are optimally allocated to the VNF chains with respect to a specific network operator's objective (e.g., maximizing the total revenue). The problem of embedding VNF chains onto the network function virtualization infrastructure is another challenging resource allocation task in the implementation of NFV.
- VNF chain embedding can be divided into two subproblems: virtual node mapping and virtual link mapping (Chowdhury et al. 2009). In VNF chain embedding, each substrate node can host a set of VNFs of different types. Since each VNF functions uniquely requires specific physical (computing, storage, or networking) resources, a VNF can only be assigned onto a substrate node that has the ability to host it. Furthermore, the VNF chain embedding stage needs to deal with dynamic networking condition changes and service description variations, which means: (1) with new services arriving and old services leaving the network, the existing VNF placement scheme may be modified to accommodate the newly coming service requests and improve resource utilization; (2) when the data rate entering an existing function chain changes, the orchestrator should notice that and trigger an adjustment algorithm to re-embed the corresponding VNF chain and reallocate the physical resources to achieve better overall network performance. Lastly, as the

intermediate stage of the NFV-RA problem, VNF chain embedding is closely related to the other two stages. It is of great importance to jointly study VNF chain composition and embedding, VNF chain embedding and VNF scheduling, and even the three stages together to obtain the optimal resource allocation solution for the NFV environment, though the problem complexity can increase seriously.

- VNF scheduling: The last stage of NFV-RA is the VNF scheduling process. Given the embedding results of VNF chains, this process tries to find a proper scheduling of VNFs' execution on each substrate node to minimize the "makespan" (the final execution time of the last VNF of the last executed network service). The results of a VNF scheduling process is closely related to the VNF chain embedding stage. The embedding process should aim to maximize the performance of the VNF scheduling stage. Therefore, it is desired to solve the VNF chain embedding and VNF-SCH problems in a coordinated way. Alternatively for simplicity, the two problems can also be solved independently where the results of VNF chain embedding are taken as the inputs to VNF scheduling (Qu et al. 2016). There are several technical challenges that need to be addressed when we consider VNF scheduling. Firstly, most of the existing proposals neglect the transmission delays of the links, how to obtain a feasible solution to the VNF scheduling problem involving the link delays remains a challenge. Moreover, currently, most existing studies focus on static scheduling algorithms in which the VNFs cannot migrate to other servers as network condition varies after performing the initial embedding and scheduling. The development of dynamic VNF scheduling algorithms remains an issue. Lastly, greedy and metaheuristic approaches exist in the literature to tackle the online VNF scheduling problem. How to formulate VNF scheduling based on mathematical optimization and solve it efficiently is also challenging.

Key Applications

It is possible and of great importance to integrate SDN and NFV in the same network to accelerate the shift toward "software-based" network services. NFV promotes the type of networking capabilities offered in service provider and cloud platforms and enables network services to be dynamically deployed among shared NFV nodes. SDN has been recognized as a key technology for the realization of a more efficient and flexible data center infrastructure. It is also a key driver in the development of edge computing and cloud computing. With the use of SDN, it becomes easier for the organizations to host applications on the most appropriate resources.

Cross-References

- [Network Function Virtualization](#)
- [Network Slicing-Enabled Green C-RAN](#)
- [Resource Allocation](#)
- [Software-Defined Wireless Networking](#)

References

- Chowdhury NMMK, Rahman MR, Boutaba R (2009) Virtual network embedding with coordinated node and link mapping. In: Proceedings of IEEE INFOCOM'09, pp 783–791
- Herrera JG, Botero JF (2016) Resource allocation in NFV: a comprehensive survey. *IEEE Netw* 13(3): 518–532
- Li J, Zhang N, Ye Q, Shi W, Zhuang W, Shen X (2017) Joint resource allocation and online virtual network embedding for 5G networks. In: Proceedings of IEEE GLOBECOM, pp 1–6
- Li J, Shi W, Ye Q, Zhuang W, Shen X, Li X (2018) Online joint VNF chain composition and embedding for 5G networks. In: Proceedings of IEEE GLOBECOM, pp 1–6
- Martini B, Paganelli F, Cappanera P, Turchi S, Castoldi P (2015) Latency-aware composition of virtual functions in 5G. In: Proceedings of Network Softwarization (Net-Soft), pp 1–6
- Qu L, Assi C, Shaban K (2016) Delay-aware scheduling and resource optimization with network function virtualization. *IEEE Netw* 64(9):3746–3758
- Xie Y, Liu Z, Wang S, Wang Y (2016) Service function chaining resource allocation: A survey. arXiv preprint arXiv:160800095

Ye Z, Cao X, Wang J, Yu H, Qiao C (2016) Joint topology design and mapping of service function chains for efficient, scalable, and reliable network functions virtualization. *IEEE Netw* 30(3):81–87

Resource Allocation in UAV-Aided Wireless Networks

Weisen Shi¹, Junling Li¹, and Ning Zhang^{2,3,4}

¹University of Waterloo, Waterloo, ON, Canada

²Peng Cheng Laboratory, Shenzhen, China

³Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

⁴College of Intelligence and Computing, Tianjin University, Tianjin, China

Synonyms

Resource allocation; UAV communication; UAV-aided wireless network

Definitions

Unmanned aerial vehicle (UAV)-aided wireless network (UWN) is designed to provide wireless network coverage to infrastructure-less or spectrum-scarce areas. Resource allocation in UWN is to assign communication, computing, or caching resources among UAVs and the connected terrestrial base stations (BSs) to improve the network performance or service quality of users.

Historical Background

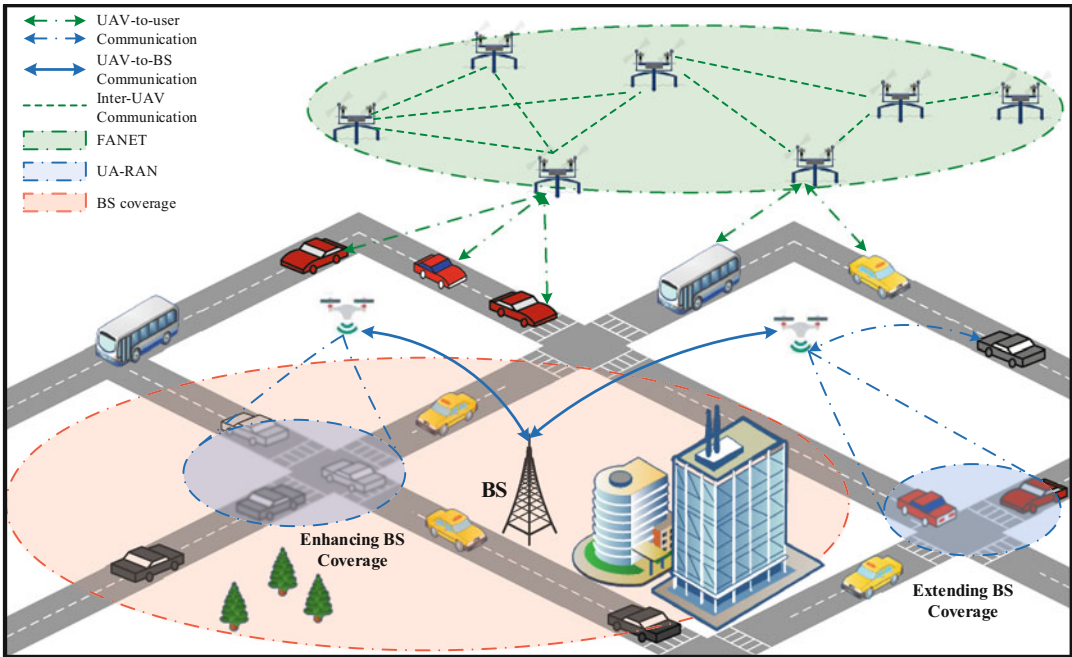
Ubiquitous connectivity is demanded by future wireless networks to support massive users with heterogeneous service requirements. However, current terrestrial wireless networks have limited radio coverage in remote or post-disaster areas, and lacks the flexibility to be re-deployed according to the dynamic variations of terrestrial user traffic. To solve the coverage and flexibility

issues, UWN is proposed by some researchers in which UAV is integrated with the wireless communication technology (Mozaffari et al. 2018). The UWN can be formed in two types, i.e., the flying ad hoc networks (FANET) composed by swarms of UAVs and the UAV-assisted radio access networks (UA-RANs) where UAVs perform as the wireless coverage extension of terrestrial BS. Specifically, the FANET is mainly used to provide temporal wireless communication services for the infrastructure-less areas (Liu and Kato 2016); while the UA-RAN is considered to enable dynamic deployment capability for future RAN to ensure the QoS for the highly variable data traffic (Bor-Yaliniz and Yanikomeroglu 2016). The two types of UWN are shown in Fig. 1. Compared with those resource allocation schemes in terrestrial networks, the communication, computing, and caching resource allocations in UWN have to jointly consider the impacts of UAV's 3D dynamic movement, i.e., the UAV static deployment (Mozaffari et al. 2017) or the UAV trajectory design (Wu et al. 2018). Compared with traditional resource allocation schemes in terrestrial networks, the impacts of UAV's 3D dynamic movement, i.e., the UAV static deployment (Mozaffari et al. 2017) or the UAV trajectory design (Wu et al. 2018), should be considered when jointly allocating the communication, computing, and caching resources.

Foundations

Compared with terrestrial wireless networks, UWN has four advantages:

- **Line-of-sight (LoS) connection:** Constrained by the two-dimensional topology, terrestrial wireless links are obstructed by vehicles or buildings frequently. As a result, most messages are transmitted through non-line-of-sight (NLoS) wireless links which can seriously decrease the quality of service (QoS) of users. Contrarily, UAVs flying in the air



Resource Allocation in UAV-Aided Wireless Networks, Fig. 1 FANET and UA-RAN

have higher probability to connect ground users via LoS links, which facilitates highly reliable communications. This advantage is further enhanced by UAV’s capability of allowing 3D adjustments of their positions to avoid obstacles between DBSs and users.

- Dynamic deployment: Different from traditional BSs which are statically fixed on dedicated locations, UAV can be dynamically deployed according to the spatial and temporal changes of ground traffic, and be assigned to different users or controllers on demands Shi et al. (2018a).
- Fully-controlled mobility: Different from connected vehicles in VANET whose mobility is controlled by drivers or autonomous vehicles themselves, the hovering positions and flying trajectories of UAV are fully controlled by the corresponding controllers. This fully-controlled mobility feature enables the dynamic deployment feature of DBSs to assist ground infrastructures.
- Communication and computation integration: The UAV holds the capability to conduct both communication and computation tasks.

Considering the dynamic deployments of UAVs, joint communication and computation resource allocation mechanisms can be implemented to further improve the performance of terrestrial networks.

To further exploit the four advantages of UWN, resource allocation in UWN can be conducted in two steps: (1) determining the optimal static deployment or trajectory design results of multiple UAVs; then (2) allocating resource for each UAV and connected terrestrial infrastructures to maximize network performance. For the first step:

- UAV static deployment: The main design objective of UAV static deployment is to optimize multiple UAVs flying altitudes, horizontal positions, and/or spatial density to achieve the maximum user coverage ratio in a given area without violating dedicated constraints, such as interference, UAV-to-ground (U2G) pathloss thresholds, etc. According to the UAV static deployment results, multiple UAVs hover on optimized positions in 3D

space and serve their associated users who require additional communication resources. Since the distribution of terrestrial users and traffic are dynamically changing in both spatial and temporal domains, the UAVs have to be re-deployed periodically based on the changes. Typically the UAV static deployment problems are formulated as the NP-hard facility location problem or similar forms of it, which are usually solved through heuristic approaches (Shi et al. 2018a).

- UAV trajectory design: Compared with the static deployment, the UAV trajectory design allows multiple UAVs fly over their associated users periodically following the designed trajectories. There are two motivations to design the trajectory of UAVs: First, for some scenarios, the number of UAVs is inadequate to cover all users through the static deployment approach. Second, the static deployment inherently lacks the user fairness guarantee. Particularly, the users located at the edge of the UAV's radio coverage can suffer from severe pathloss compared with users located at the center of the UAV radio coverage. To promote the user fairness, each UAV can periodically fly over its associated users and provide communication servers for the specific user within its scheduled interval during which the minimal UAV-to-user (U2U) pathloss can be achieved. UAV trajectory design aims to design optimal trajectories and the scheduling results for multiple UAVs to maximize network QoS without violating given constraints. UAV trajectory design problems are typically formulated as unsolvable mixed integer (U2U scheduling) nonconvex (UAV trajectory) problems, which can be solved by some simplification or iterative optimization methods, such as coordinate descent (Wu et al. 2018).

Given the static deployment or trajectory design results, the resource allocation in UWN can be categorized in two types according to the

scenarios, i.e., resource allocation in FANET and resource allocation in UA-RAN.

- Resource allocation in FANET: Resource allocation in FANET mainly focuses on how to distribute resources or tasks among the swarm of UAVs to maximize the QoS of users who access the FANET through those UAVs. Classic resource allocation schemes in traditional ad hoc networks can be implement in FANET by involving some specific features of UAVs, i.e., U2G and inter-UAV channel models (Al-Hourani et al. 2014; Dabiri et al. early access, 2018), inter-UAV interference, inter-UAV cooperation, limited CPU and battery capacities of single UAV, etc. Specifically, the resource allocation as well as the deployment and trajectory design of UAVs in FANET are usually joint optimized to maximize the user QoS gains.
- Resource allocation in UA-RAN: For UA-RAN, since UAVs are used to relay user-to-BS (U2B) communications, the resource allocation between UAVs and the corresponding terrestrial BSs are more essential than the inter-UAV resource allocation. The objective of resource allocation in UA-RAN is to optimize the resource allocation among UAVs and the terrestrial RAN to maximize the QoS of UA-RAN users without violating the QoS of terrestrial RAN users. Despite the U2G channel model and inter-UAV interference, the U2B interference and U2B link quality constraints should also be considered in the resource allocation. UA-RAN has a similar network architecture to the Cloud-RAN (C-RAN) by regarding the UAV as the “flying remote radio head (RRH).” Therefore, the RAN slicing and edge computing techniques, which are widely used in C-RAN, can be implemented in UA-RAN to promote the resource allocation. As an emerging research topic, most existing works of UA-RAN resource allocation investigate the UAV-aided caching schemes (Chen et al. 2017).

Key Applications

UWN is mainly used to extend the wireless network service to uncovered areas (remote or post-disaster regions), and provide additional communication resources for spectrum-scarce areas in terrestrial RAN's coverage (congested road/stadium or radio coverage holes) (Shi et al. 2018b). Integrated with the terrestrial networks, UWN enables the flexibility to re-deploy the access network based on the dynamic changes of users in both spatial and temporal domains. Through appropriate resource allocation, UWN can enable QoS-guaranteed ubiquitous connectivity in future wireless networks.

Cross-References

- [Ad Hoc Networks](#)
- [Resource Allocation in UAV-Aided Wireless Networks](#)
- [UAV Communication](#)

References

- Al-Hourani A, Kandeepan S, Lardner S (2014) Optimal lap altitude for maximum coverage. *IEEE Wireless Commun Lett* 3(6):569–572
- Bor-Yaliniz I, Yanikomeroglu H (2016) The new frontier in RAN heterogeneity: multi-tier drone-cells. *IEEE Commun Mag* 54(11):48–55
- Chen M, Mozaffari M, Saad W, Yin C, Debbah M, Hong CS (2017) Caching in the sky: proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience. *IEEE J Sel Areas Commun* 35(5):1046–1061
- Dabiri MT, Sadough SMS, Khalighi MA (2018) Channel modeling and parameter optimization for hovering UAV-based free-space optical links. *IEEE J Sel Areas Commun*. 36(9):2104–2113
- Liu J, Kato N (2016) A markovian analysis for explicit probabilistic stopping-based information propagation in postdisaster ad hoc mobile networks. *IEEE Trans Wirel Commun* 15(1):81–90
- Mozaffari M, Saad W, Bennis M, Debbah M (2017) Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications. *IEEE Trans Wirel Commun* 16(11):7574–7589
- Mozaffari M, Saad W, Bennis M, Nam YH, Debbah M (2018) A tutorial on UAVs for wireless networks: applications, challenges, and open problems. <https://arxiv.org/abs/1803.00680>
- Shi W, Li J, Xu W, Zhou H, Zhang N, Zhang S, Shen X (2018a) Multiple drone-cell deployment analyses and optimization in drone assisted radio access networks. *IEEE Access* 6:12518–12529
- Shi W, Zhou H, Li J, Xu W, Zhang N, Shen X (2018b) Drone assisted vehicular networks: architecture, challenges and opportunities. *IEEE Netw* 32(3):130–137
- Wu Q, Zeng Y, Zhang R (2018) Joint trajectory and communication design for multi-UAV enabled wireless networks. *IEEE Trans Wirel Commun* 17(3):2109–2121

Resource Assignment

- [Resource Allocation in Industrial IoT](#)

Resource Management

- [Interference-Aware Distributed Resource Allocation](#)
- [Resource Allocation in Industrial IoT](#)
- [Resource Allocation in NOMA-Assisted D2D Networks](#)
- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

Resource Management for Heterogeneous Networks

- [Heterogeneous Networks Through Multi-resources Deployment, Performance Enhancement for](#)

Resource Management in Macrocell—Small Cell Systems and D2D-Assisted Cellular Systems

Yao-Liang Chung

Department of Communications, Navigation and Control Engineering, National Taiwan Ocean University, Keelung, Taiwan

Synonyms

Synonym 1. Macrocell—small cell systems; Heterogeneous cellular networks

Synonym 2. Device-to-device (D2D)-assisted cellular systems; D2D communication in cellular systems

Definitions

Definition 1 (Macrocell—Small Cell System).

A two-layer cellular network architecture in which the area covered by a macrocell base station (MBS) has multiple small cell base stations (SBSs) deployed within it (i.e., the areas covered by the SBSs are incorporated into the larger overall MBS coverage area). The term “heterogeneous cellular networks” is also used to refer to such concurrent operation of MBSs and SBSs.

Definition 2 (Device-to-device (D2D)-Assisted Cellular System).

A cellular network architecture that utilizes so-called D2D communication, that is, communication in which data can be sent directly (i.e., without first being sent to a base station (BS)) by each device to its destination device, thus allowing for the replacement of the cellular mode (through both uplinking and downlinking) for transmissions sent over short distances. Such a network architecture is also referred to as “D2D communication in a cellular system.”

Introduction

The study of cellular networks, which are an important type of wireless network, covers not just technological dimensions but also social and economic dimensions. Since the start of the transition from second- to third-generation cellular networks, communication standards have been regulated under the authority of the Third Generation Partnership Project (3GPP), an international collaboration among all the relevant companies and organizations. In a cellular network, base stations (BSs) play an important role as they enable devices connected to the system (BS) to communicate with each other. In first-, second-, and third-generation cellular networks, the radio access network (RAN) side can only utilize macrocell base stations (MBSs) to provide network services to their users. However, the rapid growth in the number of mobile phone users, increasing shipment volumes of smart handheld devices, and emerging application services have aggravated the problems associated with the surge in global mobile data flow (Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2016) and the imbalance between supply and demand. Telecom operators can no longer rely on a single technology (i.e., MBSs) to handle these problems. Furthermore, users seeking to operate devices indoors, at macrocell edges, or in certain locations such as caves, basements, elevators, etc. will occasionally encounter poor signal conditions or even signal dead zones (i.e., areas in which they are not able to get a signal from an MBS). Both of the aforementioned facts have together led to the development of heterogeneous access technologies (e.g., small cell base stations (SBSs) and device-to-device (D2D) communication) for flow offloading and coverage expansion in local area networks. Fourth-generation cellular networks went a step further by integrating a variety of heterogeneous access technologies (including Wi-Fi, SBSs, and D2D) to raise overall network performance. These technologies focused on improving factors such as spectral efficiency, energy efficiency,

and network coverage, so as to create a good mobile network usage environment for users. Going forward, fifth- and later-generation cellular networks will be shifting toward the use of multi-network heterogeneous network architectures that are even more advanced and efficient (Gupta and Jha 2015).

It is worth noting that SBS and D2D technologies, as well as the application prospects that they offer, are gaining widespread attention. These two technologies have opened up new opportunities with respect to local area network layouts for telecom operators and businesses dealing with equipment, chips, and software, and these businesses have made the two technologies the focal points of their research on fifth-generation cellular networks. However, regardless of the ways in which heterogeneous networks are further developed and deployed, efficient resource management (which covers key aspects including time, frequency, space, and power) still remains an important means and step in determining if a system's performance can meet its various objectives and demands. This entry will provide an introduction to macrocell—small cell systems and D2D-assisted cellular systems, which are created through the application of SBS and D2D technologies to macrocell networks. A comprehensive and up-to-date survey on resource management for such systems will also be given, so as to facilitate a discussion of potential commercial applications and future challenges.

Resource Management in Macrocell—Small Cell Systems

Background

SBS technologies have become a part of the discussion on standardization since 3GPP Release 8, and the key points of this discussion have included potential issues such as interference issues among SBSs, interference between MBSs and SBSs, and SBS deployment optimization in user-intensive areas. In comparison to MBSs, SBSs are access nodes that are lower in cost and use less energy but also have smaller coverage ranges (Small Cell Forum 2017). Because the

different types of small cells differ in terms of average transmission power and cell size, they are frequently classified into various subcategories including microcells, picocells, and femtocells, with various scholars having noted that investigations aimed at moving toward the use of smaller and smaller cells should constitute a major thread of future research (Fehske et al. 2014).

SBSs were originally developed to address poor indoor signals and macrocell edge signals, as well as signal dead zones. But with the emergence of fourth-generation cellular network architectures, which differ greatly from the ones used by previous generations, SBSs have become a deployment focus alongside MBSs. Moreover, the rapid growth of mobile data services has led to SBSs being utilized to offload mobile network flows and provide additional value-added services. In order to reduce operating costs and facilitate the installation of SBSs by users, telecom operators also require SBSs to be able to form self-organizing networks, such that they can perform self-configuration and optimization effectively. It is expected that, so long as the interferences among SBSs and between MBSs and SBSs can be sufficiently controlled, the communication quality for all application services will be considerably enhanced by the efficient deployment of SBSs within the coverage areas covered by MBSs.

Literature Survey

It should be noted, first, that a considerable amount of the existing research regarding the management of cellular network resources has considered only the involvement of MBSs in such networks, with a large number of studies having provided comprehensive investigations of how to improve system performance via the unique contributions of various approaches. For example, an efficient cross-layer packet scheduling and resource management scheme was proposed as a means of maximizing system capacity by Jeong et al. (2006), while fast algorithms aimed at computing the optimal allocation of resources for overall system utility maximization were proposed by Madan et al.

(2010). In still another approach, Rodrigues and Casadevall (2011) adopted two viewpoints in studying the trade-offs made between user fairness and resource utilization, whereas studies by Ng et al. were dedicated to investigating optimal resource allocation to ensure energy-efficient transmission in the context of both single-cell environments (Ng et al. 2012a) and multicell environments (Ng et al. 2012b). In a subsequent study, Ng et al. explored the effects of utilizing another type of MBS for data transmissions, specifically, an MBS that harvests hybrid energy (Ng et al. 2013). Meanwhile, due to ongoing advances in cellular technologies, it is now possible to simultaneously use more than one component carrier (CC) within a BS for transmissions. As such, a number of recent studies have consisted of investigations of systems involving MBSs that simultaneously utilize multiple CCs. For example, a study by Chung (2015a) proposed a rate-and-power control transmission scheme aimed at minimizing the amount of energy consumed by the transceivers of an MBS making use of two CCs at once while still ensuring that users would receive the required service quality and fairness. In a subsequent paper (Chung 2015b), the same author considered the same system in order to construct a framework that was then used to comprehensively examine potential means of maximizing transmission efficiency through both the maximization of system capacity and the minimization of the total amount of energy consumed. Chung followed that study by yet another in which he proposed a more flexible scheme aimed at both efficiently minimizing the total power consumed by the transceivers of an MBS when said MBS is simultaneously making use of two or more CCs and ensuring that the necessary service quality and fairness are still provided to users (Chung 2016a). The author then followed that study by still another proposing an efficient packet scheduling algorithm that could be used to increase transmission efficiency when both real-time traffic and non-real-time traffic are being transmitted (Chung 2016b).

Next, it is important to note that when multiple SBSs and MBSs are both present and used simultaneously in a given system, a range of complex resource management issues are consequently created. As such, various investigations focusing their attention on these issues have also been conducted. For example, members of the aforementioned 3GPP organization have conducted a range of discussions regarding potential approaches to minimizing the power consumed by evolved universal terrestrial radio access networks (3GPP TR. 36.927 V10.1.0 2011). Relatedly, a scheme proposed by Claussen et al. (2010) would utilize the detection of user activity to enhance energy efficiency, while a study conducted by Wang and Shen (2010) investigated the energy efficiency of individual cells (as measured in bits per Joule) as well as the energy efficiency of specific coverage areas (as measured in energy efficiency per unit area). The following year, 2011, a study was conducted by Auer et al. (2011) in which the different levels of energy used by different types of BSs were quantified, while a subsequent study by Aleksic et al. (2013) proposed a hybrid model to be used for assessing the energy efficiency of converged wireless/optical access networks. In 2015, 2 years after the publication of the study by Aleksic et al., Huang et al. (2015) proposed both distributed and centralized optimization methods aimed at addressing problems encountered in the optimization of small cell cluster coverage, while in 2016, the possibility of imposing quality-of-service requirements and interference mitigation was investigated by Wang et al. as a potential means of enhancing transmission efficiency (Wang et al. 2016). Meanwhile, a number of transmission algorithms aimed at ensuring that data rate needs would be met even under various SBS deployment strategies and even while minimizing the total power consumed by the transceivers of all the BSs were proposed by Chung (2015c, 2017a). In addition, a 2014 study by Mekikis et al. (2014) proposed a method for determining the lowest cost of network deployment with random deployment, while in a still more recent study, a continuous-time Markov chain model that would allow the survivability

of networks subjected to extensive failures (such as failures stemming from natural disasters and security attacks, as well as common mode software and hardware failures) to be quantified was proposed by Xie et al. (2017). Meanwhile, systems involving the simultaneous activity of multiple mobile network operators have been the focus of other recent studies. For example, Antonopoulos et al. (2015) proposed a framework for providing enhanced throughput and energy efficiency in such systems via intra-operator spectrum sharing, while Bousia et al. proposed an auction-based switching off and offloading algorithm that could allow energy conservation and cost reductions by encouraging mobile network operators to switch off any redundant MBSs and offload their traffic (Bousia et al. 2016). Relatedly, a study conducted by Oikonomakou et al. (2017) investigated the energy and cost efficiency of a network effectively comprised of the cooperative actions of multiple mobile network operators. More recently still, other studies have given close attention to the problem of the signal dead zones that can sometimes be found in such heterogeneous cellular networks. For example, Chung (2017b) introduced a coverage algorithm aimed at providing the greatest possible decrease in the total power utilized by all of the transceivers of the BSs in a system while at the same time ensuring that comprehensive wireless signal coverage would be provided to users under a variety of scenarios; however, the algorithm's lone disadvantage was that it was not always able to fulfill the transmission rate requirements of users. Chung (2017c) then proposed a modified coverage algorithm to mitigate this disadvantage and ensure that the respective transmission rate requirements of all users can be met.

Commercial Applications

Currently, SBSs are primarily used to strengthen indoor signals and prevent signal dead zones, and relevant businesses are also hoping to utilize SBSs to operate value-added application services in specific regions. By being integrated with application-end equipment or devices, SBSs can

then be utilized in smart grids (Bu and Yu 2014), smart homes (Khan et al. 2015), and smart cities (Zhou et al. 2016), thus allowing for economic benefits to be generated.

Challenges

The deployment of SBSs in a high-density manner is viewed as one particularly promising approach to addressing the increasing demand of high capacity for next-generation cellular networks, with this approach being expected to be studied extensively in future research. However, one key point to address would be the various issues that arise from such deployments, which include signal interference, frequency band use, deployment strategies, backhaul/fronthaul high traffic load designs, and operation management.

Resource Management in D2D-Assisted Cellular Systems

Background

As for D2D technologies, they became a part of the discussion on standardization in 3GPP Release 12, and the key points of this discussion have included potential issues such as proximity-based services, device search mechanisms, resource management, interference control, converged interconnection, group communication, and security and privacy. In fact, long before the introduction of 3GPP Release 12, the development of D2D technologies (such as Bluetooth) had been ongoing for some time, yet there had been a lack of results with respect to the development of key applications and, correspondingly, the commercial adoption of such technologies. However, in recent years, the emergence of a multitude of smart mobile devices and mobile services and applications has resulted in a surge in mobile data flow, the growth of community networks, and the increasing application of the Internet of Things (IoT), all of which have attracted renewed industry and academic interest in D2D technologies.

In traditional cellular networks, the RAN side can only establish communication links with the support of MBSs. But with the application of

D2D technologies, mobile devices within a certain range can perform direct data transfers without the support of MBSs, and at this point, every device in the network will possess the ability to transmit, receive, and relay data. Through effective resource management, D2D technologies can provide advantages such as offloading for MBSs, the reduction of transmission delays, the raising of data rates, spectral efficiency, energy efficiency, and the expansion of network coverage.

Literature Survey

D2D communication can be divided into four categories according to the different kinds of frequency bands (specifically, licensed bands and unlicensed bands) that past studies have used for D2D communication: (1) D2D communication for which only licensed bands are used (such as that described in Asadi et al. (2014a) and Doppler et al. (2009); Janis et al. (2009); Zulhasnine et al. (2010); Feng et al. (2013); Wen et al. (2013); Wang et al. (2012); Akkarajitsakul et al. (2012); Doppler et al. (2010); Chung and Lin (2014); Gao et al. (2017); Giatsoglou et al. (2017); Chen et al. (2017); Datsika et al. (2016)), (2) D2D communication for which only unlicensed bands are used (such as that described in Asadi et al. (2014a) and Asadi and Mancuso (2013); Golrezaei et al. (2014); Dimitriou et al. (2013); Mastorakis et al. (2013); Bourdena et al. (2014); Antonopoulos et al. (2014); Datsika et al. (2016, 2017)), (3) D2D communication for which either licensed or unlicensed bands are used in a selective manner (such as that described in Asadi et al. (2014b, 2015)), and (4) D2D communication for which both licensed and unlicensed bands are used simultaneously (such as that described in Chung (2017d, e)). With regard to the first two categories, Asadi et al. (2014a) reported a D2D communication survey and overview that showed that D2D connections can effectively mitigate the transmission delays and traffic loads faced by MBSs, with those D2D connections also resulting at the same time in enhanced system throughput and better energy usage efficiency (effects which in turn result in decreases in the amount of overall power utilized by a network). Relatedly,

the amount of clearly established research results regarding the first two categories is greater than the amount of such research results regarding the third and fourth categories; as such, there is substantially more research progress that can still be accomplished with respect to the latter two categories.

According to a study by Doppler et al. (2009), one way of potentially reducing the effects of interference to a substantial degree would consist of imposing limitations on the distances of transmissions and the transmission power consumed by D2D pairs. Meanwhile, resource allocation schemes that would potentially cause interference levels to be reduced have also been proposed by Janis et al. (2009) and Zulhasnine et al. (2010), although one limitation of both of those schemes is that they can be employed only for D2D pairs themselves. More recently, still other potentially beneficial resource allocation schemes have been presented by other relevant studies, including those by Feng et al. (2013) and Wen et al. (2013). That said, while a variety of D2D communication studies have been produced, it should be noted that system throughput was the primary focus of much of that research. However, given that mobile devices typically have a limited amount of battery energy available to them, and because energy efficiency is increasingly seen as a key consideration in the design of both mobile devices and network schemes, the efficient use of power and various related subjects has recently been given greater research attention. For example, in a study concerning uniformly distributed D2D networks, Wang et al. (2012) suggested the use of reverse iterative combinatorial auction games as one potential means of providing energy-efficient resource allocation, while in the previously mentioned study conducted by Wen et al. (2013), the question of how to improve D2D cluster energy efficiency was explored. Still other studies have considered the opportunistic switching of communication modes as another potential method for addressing the same issue (Akkarajitsakul et al. 2012; Doppler et al. 2010), while more recently, Chung and Lin (2014)

introduced a transmission algorithm intended to reduce overall power consumption while also making sure that all users would still have their transmission rate requirements met. Meanwhile, Gao et al. (2017) conducted an analysis of how energy efficiency changes in D2D-assisted cellular networks alter spectrum efficiency, and vice versa, while new techniques for potentially improving system performance in D2D communication have been explored in a number of other recent studies. For example, Giatoglou et al. (2017) proposed a device caching approach intended to facilitate the exchange of content via millimeter-wave D2D communication, while an approach based on the use of D2D communications to leverage any substantial crowd of devices located at a network's edge as a means of providing energy-efficient collaborative task execution was proposed by Chen et al. (2017). It should be reiterated, incidentally, that the D2D-related studies just described all considered the issue of transmission over licensed bands.

That said, the subject of D2D communication over unlicensed bands has recently become a subject of greater interest among researchers due to the simple fact that the majority of newer mobile devices use at least two wireless interfaces when operating. For example, a protocol focusing on the topic of implementation was proposed in a paper by Asadi and Mancuso (2013), while the use of localized D2D communications as a means of increasing the throughput of video files transmitted via cellular communication systems was suggested by Golrezaei et al. (2014). The year before, in 2013, Dimitriou et al. (2013) published a study in which a device-focused methodology was used to analyze the design, development, and experimental evaluation of a scheme for heterogeneous wireless networks based on the use of delay-bounded energy-conscious bandwidth allocation. In the same year, Mastorakis et al. (2013) published a study proposing an energy-efficient routing scheme intended to enhance the energy efficiency and data transmission reliability communication devices possessing heterogeneous radio spectrum availability. The following year, Bourdena et

al. (2014) expanded upon that earlier study by Mastorakis et al. (2013). It should be noted, meanwhile, that whenever multiple devices attempt to engage in D2D communications over unlicensed bands, channel access issues will inevitably be raised, particularly in those situations wherein all the devices involved have the same aim with regard to content dissemination. The need to identify effective schemes for providing medium access control (MAC) is underscored by the occurrence of such situations, and in fact, the relevant issues have been explored and addressed in several recent studies. For example, Antonopoulos et al. (2014) proposed a pair of game theoretic MAC schemes aimed at providing rapid and energy-efficient content dissemination, while Datsika et al. (2017) proposed a network coding-based MAC approach for data exchange via D2D transmissions. In a slightly earlier study, meanwhile, an overview of D2D communication and MAC design challenges involving energy consumption issues and social behaviors was also provided by Datsika et al. (2016).

One key issue impacting the use of either licensed bands only or unlicensed bands only to complete D2D transmissions is that both approaches provide differing advantages and disadvantages with respect to the resulting implementation complexity, spectral efficiency, and interference control (for more details, see Asadi et al. (2014a)). Therefore, the overall performance of the system in question will suffer if just one or the other approach is used exclusively. With that in mind, the selective use of both licensed and unlicensed bands was examined by Asadi et al. in two of their recent papers (Asadi et al. 2014b, 2015). Specifically, they suggested a heuristic solution called "Floating Band D2D" as one potential means of making better use of the potential benefits of utilizing both licensed and unlicensed bands. That is, their proposed solution was intended to guide a network such that either licensed or unlicensed bands would be used appropriately for a given D2D data transmission. More recently, Chung (2017d) published a study investigating the full and simultaneous use of both licensed

and unlicensed bands for D2D communication wherein the goal of providing users with more optimized average transmission rates was effectively accomplished. In yet another study, meanwhile, the same author (Chung 2017e) proposed a transmission algorithm based on the deployment of novel system configurations as a means of providing enhanced utility in using both the licensed and unlicensed bands; the results indicated that the proposed algorithm could not only provide all connected users with acceptable transmission rates but that it could also improve system performance in terms of both connection efficiency and energy efficiency to a significant degree.

Commercial Applications

In cellular systems, the use of D2D communication is directly and practically relevant to a wide variety of information and communication technology applications, including the IoT (Xu et al. 2014), Internet of Vehicles (Ferreira et al. 2014), wireless sensor networks (Marchenko et al. 2014), and wearable devices/sensors for healthcare (Lin et al. 2016). Some potential examples are discussed in the aforementioned work (Chung 2017d) by Chung. The application of D2D communication in the abovementioned fields can lead to the development of application services that carry high social and economic value.

Challenges

At present, the subject of efficient D2D-assisted transmission involving simultaneous and completely flexible use of both licensed and unlicensed bands in cellular systems has yet to be comprehensively researched. Given the substantial relevance of this topic to next-generation cellular networks, further research is certainly required with regard to the question of how both power and radio resources can be efficiently employed in D2D-assisted cellular systems utilizing both licensed and unlicensed bands. Relatedly, research regarding the integration of SBS and D2D technologies in a single system as means of further raising network performance is also gaining increasing momentum.

Cross-References

- [Heterogeneous Small Cell Networks](#)
- [Relay-Involved Device-to-Device Communications in LTE Networks](#)

References

- 3GPP TR. 36.927 V10.1.0 (2011) Potential solutions for energy saving for E-UTRAN (Release 10). Available online: <http://www.qtc.jp/3GPP/Specs/36927-a10.pdf>. Accessed 10 Mar 2017
- Akkarajitsakul K, Phunchongharn P, Hossain E, Bhargava V (2012) Mode selection for energy-efficient d2d communications in lte-advanced networks: a coalitional game approach. In: Proceedings of IEEE ICCS, Singapore
- Aleksic S, Deruyck M, Vereecken W, Joseph W, Pickavet M, Martens L (2013) Energy efficiency of femtocell deployment in combined wireless/optical access networks. *Comput Netw* 57:1217–1233
- Antonopoulos A, Kartsakli E, Verikoukis C (2014) Game theoretic D2D content dissemination in 4G cellular networks. *IEEE Commun Mag* 52:125–132
- Antonopoulos A, Kartsakli E, Bousia A, Alonso L, Verikoukis C (2015) Energy-efficient infrastructure sharing in multi-operator mobile networks. *IEEE Commun Mag* 53:242–249
- Asadi A, Mancuso V (2013) WiFi direct and LTE D2D in action. In: Proceedings of IEEE wireless days, Valencia
- Asadi A, Wang Q, Mancuso V (2014a) A survey on device-to-device communication in cellular networks. *IEEE Commun Surveys Tutor* 16:1801–1819
- Asadi A, Jacko P, Mancuso V (2014b) Modeling D2D communications with LTE and WiFi. *ACM SIGMETRICS Perform Eval Rev* 42:55–57
- Asadi A, Mancuso V, Jacko P (2015) Floating band D2D: exploring and exploiting the potentials of adaptive D2D-enabled networks. In: Proceedings of IEEE WoW-MoM, Boston
- Auer G, Giannini V, Desset C, Godor I, Skillermark P, Olsson M, Imran MA, Sabella D, Gonzalez MJ, Blume O, Fehske A (2011) How much energy is needed to run a wireless network? *IEEE Wirel Commun* 18:40–49
- Bourdena A, Mavromoustakis CX, Kormentzas G, Pallis E, Mastorakis G, Yassein MB (2014) A resource intensive traffic-aware scheme using energy-aware routing in cognitive radio networks. *Futur Gener Comput Syst* 39:16–28
- Bousia A, Kartsakli E, Antonopoulos A, Alonso L, Verikoukis C (2016) Multiobjective auction-based switching-off scheme in heterogeneous networks: to bid or not to bid? *IEEE Trans Veh Technol* 65:9168–9180
- Bu S, Yu FR (2014) Green cognitive mobile networks with small cells for multimedia communications in the smart grid environment. *IEEE Trans Veh Technol* 63:2115–2126

- Chen X, Pu L, Gao L, Wu W, Wu D (2017) Exploiting massive D2D collaboration for energy-efficient mobile edge computing. *IEEE Wirel Commun* 24:64–71
- Chung Y-L (2015a) Rate-and-power control based energy-saving transmissions in OFDMA-based multicarrier base station. *IEEE Syst J* 9:578–584
- Chung Y-L (2015b) Energy-efficient transmissions for green base stations with a novel carrier activation algorithm: a system-level perspective. *IEEE Syst J* 9:1252–1263
- Chung Y-L (2015c) An energy-saving small-cell zooming scheme for two-tier hybrid cellular networks. In: *Proceedings of IEEE ICOIN, Siem Reap*
- Chung Y-L (2016a) A novel power-saving transmission scheme for multiple-component-carrier cellular systems. *Energies* 9:265
- Chung Y-L (2016b) A novel algorithm for efficient downlink packet scheduling for multiple-component-carrier cellular systems. *Energies* 9:950
- Chung Y-L (2017a) Energy-saving transmission for green macrocell—small cell systems: a system-level perspective. *IEEE Syst J* 11:706–716
- Chung Y-L (2017b) An energy-efficient coverage algorithm for macrocell—small cell network systems. *Energies* 10:1319
- Chung Y-L (2017c) An improved algorithm for transmission in heterogeneous cellular networks. In: *Proceedings of IEEE ICDIM, Fukuoka*
- Chung Y-L (2017d) ETLU: enabling efficient simultaneous use of licensed and unlicensed bands for D2D-assisted mobile users. *IEEE Syst J*. <https://doi.org/10.1109/JSYST.2017.2651953>
- Chung Y-L (2017e) Energy-efficient use of licensed and unlicensed bands in D2D-assisted cellular network systems. *Energies* 10:1893
- Chung Y-L, Lin M-Y (2014) A power-saving resource allocation algorithm for D2D-assisted cellular networks. In: *Proceedings of IEEE INTECH, Luton*
- Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White paper. Available online: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visualnetworking-index-vni/mobile-white-paper-c11-520862.html>. Accessed 15 Apr 2017
- Claussen H, Ashraf I, Ho LTW (2010) Dynamic idle mode procedures for femtocells. *Bell Syst Tech J* 15:95–116
- Datsika E, Antonopoulos A, Zorba N, Verikoukis C (2016) Green cooperative device-to-device communication: a social-aware perspective. *IEEE Access* 4:3697–3707
- Datsika E, Antonopoulos A, Zorba N, Verikoukis C (2017) Cross-network performance analysis of network coding aided cooperative outband D2D communications. *IEEE Trans Wirel Commun* 16:3176–3188
- Dimitriou CD, Mavromoustakis CX, Mastorakis G, Pallis E (2013) On the performance response of delay-bounded energy-aware bandwidth allocation scheme in wireless networks. In: *Proceedings of IEEE ICC workshops, Budapest*
- Doppler K, Rinne M, Wijting C, Ribeiro C, Hugl K (2009) Device-to-device communication as an underlay to lte-advanced networks. *IEEE Commun Mag* 47:42–49
- Doppler K, Yu C-H, Ribeiro C, Janis P (2010) Mode selection for device-to-device communication underlying an lte-advanced network. In: *Proceedings of IEEE WCNC, Sydney*
- Fehske AJ, Viering I, Voigt J, Sartori C, Redana S, Fettweis GP (2014) Small-cell self-organizing wireless networks. *Proc IEEE* 102:334–350
- Feng D, Lu L, Yuan-Wu Y, Li G, Feng G, Li S (2013) Device-to-device communications underlying cellular networks. *IEEE Trans Commun* 61:3541–3551
- Ferreira JC, Monteiro V, Afonso JL (2014) Vehicle-to-everything application (V2Anything App) for electric vehicles. *IEEE Trans Ind Informat* 10:1927–1937
- Gao H, Wang M, Lv T (2017) Energy efficiency and spectrum efficiency tradeoff in the D2D-enabled HetNet. *IEEE Trans Veh Technol* 66:10583–10587
- Giatsoglou N, Ntontin K, Kartsakli E, Antonopoulos A, Verikoukis C (2017) D2D-aware device caching in mmWave-cellular networks. *IEEE J Sel Areas Commun* 35:2025–2037
- Golrezaei N, Mansourifard P, Molisch AF, Dimakis AG (2014) Base-station assisted device-to-device communications for high-throughput wireless video networks. *IEEE Trans Wirel Commun* 13:3665–3676
- Gupta A, Jha RK (2015) A survey of 5G network: architecture and emerging technologies. *IEEE Access* 3:1206–1232
- Huang L, Zhou Y, Wang Y, Han X, Shi J, Chen X (2015) Advanced coverage optimization techniques for small cell clusters. *China Commun* 12:111–122
- Janis P, Koivunen V, Ribeiro C, Korhonen J, Doppler K, Hugl K (2009) Interference-aware resource allocation for device-to-device radio underlying cellular networks. In: *Proceedings of IEEE VTC Spring, Barcelona*
- Jeong SS, Jeong DG, Jeon WS (2006) Cross-layer design of packet scheduling and resource allocation in OFDMA wireless multimedia networks. In: *Proceedings of IEEE VTC 2006 Spring, Melbourne*
- Khan MA, Leng S, Huang X, Yin J, Fan B (2015) Backhaul aggregation for smart homes in heterogeneous wireless networks. In: *Proceedings of IEEE CIT/IUCC/DASC/PICOM, Liverpool*
- Lin F, Wang A, Zhuang Y, Tomita MR, Xu W (2016) Smart insole: a wearable sensor device for unobtrusive gait monitoring in daily life. *IEEE Trans Ind Informat* 12:2281–2291
- Madan R, Boyd S, Lall S (2010) Fast algorithms for resource allocation in wireless cellular networks. *IEEE/ACM Trans Netw* 18:973–984
- Marchenko N, Andre T, Brandner G, Masood W, Bettstetter C (2014) An experimental study of selective cooperative relaying in industrial wireless sensor networks. *IEEE Trans Ind Informat* 10:1806–1816
- Mastorakis G, Mavromoustakis CX, Bourdena A, Pallis E (2013) An energy-efficient routing scheme using backward traffic difference estimation in cognitive radio networks. In: *Proceedings of IEEE WoWMoM, Madrid*
- Mekikis P-V, Kartsakli E, Antonopoulos A, Lalos AS, Alonso L, Verikoukis C (2014) Two-tier cellular

- random network planning for minimum deployment cost. In: Proceedings of IEEE ICC, Sydney
- Ng DWK, Lo ES, Schober R (2012a) Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas. *IEEE Trans Wirel Commun* 11:3292–3304
- Ng DWK, Lo ES, Schober R (2012b) Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity. *IEEE Trans Wirel Commun* 11:3618–3631
- Ng DWK, Lo ES, Schober R (2013) Energy-efficient resource allocation in OFDMA systems with hybrid energy harvesting base station. *IEEE Trans Wirel Commun* 12:3412–3427
- Oikonomakou M, Antonopoulos A, Alonso L, Verikoukis C (2017) Evaluating cost allocation imposed by cooperative switching off in multi-operator shared HetNets. *IEEE Trans Veh Technol* 66:11352–11365
- Rodrigues EB, Casadevall F (2011) Control of the trade-off between resource efficiency and user fairness in wireless networks using utility based adaptive resource allocation. *IEEE Commun Mag* 49:90–98
- Small Cell Forum. Available online: <http://www.smallcellforum.org>. Accessed 8 June 2017
- Wang W, Shen G (2010) Energy efficiency of heterogeneous cellular network. In: Proceedings of IEEE VTC 2000 Fall, Ottawa
- Wang F, Xu C, Song L, Han Z, Zhang B (2012) Energy-efficient radio resource and power allocation for device-to-device communication underlying cellular networks. In: Proceedings of IEEE WCSP, Huangshan
- Wang Y, Dai X, Wang JM, Bensaou B (2016) Iterative greedy algorithms for energy efficient LTE small cell networks. In: Proceedings of IEEE WCNC, Doha
- Wen S, Zhu X, Lin Z, Zhang X, Yang D (2013) Energy efficient power allocation schemes for device-to-device(d2d) communication. In: Proceedings of IEEE VTC Fall, Las Vegas
- Xie L, Heegaard PE, Jiang Y (2017) Survivability analysis of a two-tier infrastructure-based wireless network. *Comput Netw* 128:28–40
- Xu LD, He W, Li S (2014) Internet of things in industries: a survey. *IEEE Trans Ind Informat* 10:2233–2243
- Zhou L, Wei L, Sheng Z, Hu X, Zhao H, Wei J, Leung VCM (2016) Green cell planning and deployment for small cell networks in smart cities. *Ad Hoc Netw* 43:30–42
- Zulhasnine M, Huang C, Srinivasan A (2010) Efficient resource allocation for device-to-device communication underlying lte network. In: Proceedings of IEEE WiMob, Niagara Falls

Resource Planning

- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

Resource Scheduling

- [Resource Allocation in Industrial IoT](#)

Resource Scheduling in Virtualized Wireless Networks

Xianfu Chen

VTI Technical Research Centre of Finland Ltd,
Oulu, Finland

Synonyms

Elastic wireless resource management; Wireless resource virtualization

Definition

Resource scheduling refers to the sequences of decision-makings from an entity ensure the accomplishment of tasks arriving to the network based on the resource availabilities. In virtualized wireless networks, the resource can be not only network slices with specifically designed functionalities but isolated frequency spectrum.

History

The proliferation of mobile devices and data-intensive applications has caused a significant demand of wireless network services to grow at a rapid rate. The global mobile data traffic is estimated to increase eightfold between 2015 and 2020 (Cisco 2016). To keep pace with the enormous demands, the current spectrum efficiency has to be extended by at least a factor of 10. Network densification has been considered as the dominant theme towards future wireless networking (Andrews et al. 2012). That is, the

cells tend to be smaller, which indicates that the traditional border among different radio access technologies will disappear. On one hand, an advanced dense network infrastructure can certainly achieve seamless mobility, much higher network capacity and spectrum and energy efficiency due to shorter transmission range and less number of mobile users per cell site. On the other hand, the mobile data traffic evolves from being voice-dominated to being data-dominated. The revenue per unit-data for wireless network operators is declining at an unhealthy rate, whereas the costs are increasing considerably with the heavy budgetary burden of upgrading network infrastructures. All these trends are pushing the wireless network operators to find a better way to provide wireless services.

Due to the potential performance gains and the inherent cost-saving benefits by utilizing economics of scale and avoiding duplicated investment on the network infrastructure, radio access network sharing among the wireless network operators has attracted extensive attention and emerges as a promising mechanism to control both the capital expenditure and the operational expenditure (Liang and Yu 2015). More specifically, if a wireless service subscriber can freely connect to a base station of the wireless network operators and the base stations of one wireless network operator can reuse the spectrum of other wireless network operators, the network capacity can be significantly enhanced to accommodate more traffic generated from the subscribers. The worldwide combined capital expenditure and operational expenditure savings can be up to \$60 billion through radio access network sharing over a period of 5 years (Cellular News 2009).

However, the radio access network sharing activities are mostly based on long-term business agreements between the wireless network operators, and most existing radio access network sharing solutions have the drawbacks of orchestrating both data and control planes among the wireless network operators, customizing wireless services, and capability of adapting to dynamic network statistics in real world. Network virtualization facilitates radio access

network sharing, with which the traditional single ownerships of network infrastructure and spectrum resources can be separated from the wireless services. Consequently, the same physical network infrastructure is able to host multiple wireless service providers. The radio access in the virtualized wireless networks can be coordinated by a software-defined networking framework (Gudipati et al. 2013). Software-defined networking simplifies network management by decoupling the control plane from the data plane and is a natural platform for realizing network virtualization in radio access networks.

The recent advances in software-defined networking, cloud computing, as well as edge computing (Satyanarayanan 2017) provide technical enablers for a software-defined wireless network. The origin of the software-defined networking concept can be traced back to the 1990s. The success of software-defined networking comes from the systematic abstraction of complex networking problems in the Internet, which turns distributed networking problems into a logically centralized one. Software-defined networking gives rise to fundamental new thinking on the design of wireless networks. The development of a software-defined wireless network is highly connected to cloud computing which makes large-scale logical centralized control solutions feasible. Yet the traditional cloud computing architecture may have a problem in satisfying the strict latency requirements lying in control plane functioning. It is hence reasonable to move the logical centralized control close to the edge in wireless networks. Edge computing could fill this gap for a better architecture design of the software-defined wireless networks. Edge computing is a variant of the cloud computing concept but uses the computation resources and storages at the edge of a wireless network for a sustainable amount of communication, storage, control, and configuration. In a wireless network, edge computing can be utilized for the control and joint signal processing at the radio access network to serve densely deployed cells, while cloud computing can be used in core networks for packet processing and forwarding. The integration of edge

computing and cloud computing brings an end-to-end software-defined solution for the wireless networks.

Foundations

The decoupling of the control and data planes in a software-defined wireless network makes it a must to take into account different timescales in the scheduling of wireless radio resource. Wireless resource needs to carefully be split locally or remotely in order to match the new network architecture. The radio access network sharing among the wireless service providers add another level of challenges to wireless resource scheduling. The radio access network slicing is not simply the spectrum slicing but means the wireless resource isolation. Large-scale learning is an enabling technology that allows the wireless service providers to dig out the subscriber- and the service-level awareness in traffic over the widespread wireless networks and is a promising facilitator of real-time conversations between the wireless service providers and the virtualized wireless networks to influence the quality-of-service as well as the quality-of-experience of service subscribers, adapting to the network dynamics and the service usage.

The problem of wireless resource scheduling in a virtualized radio access network involving multiple wireless service providers is in general formulated as a multi-agent Markov decision process. In a large-scale multi-agent network setting, having complete environmental information over the time horizon is extremely challenging from the perspective of a wireless service provider. Alternatively, a wireless service provider makes partial observations of the environment, from which the abstract environmental information can be explored. Following this logic, we are able to implement the large-scale learning for wireless resource scheduling, which is then simplified to a series of independent single-agent small-scale learning.

Cross-References

- [Network Slicing-Enabled Green C-RAN](#)
- [Spectrum Resource Allocation in Cognitive Radio Networks](#)

References

- Andrews JG, Claussen H, Dohler M, Rangan S, Reed MC (2012) Femtocells: past, present, and future. *IEEE J Sel Areas Commun* 30(3):497–508
- Cellular News (2009) Active ran sharing could save operators \$60 billion. [Online] Available at: <http://www.cellular-news.com/story/36831.php>. [Haettu 21 May 2018]
- Cisco (2016) Cisco visual networking index: forecast and methodology, 2015–2020. White paper from Cisco
- Gudipati A, Perry D, Li LE, Katti S (2013) SoftRAN: software defined radio access network. ACM, Hong Kong
- Liang C, Yu FR (2015) Wireless network virtualization: a survey, some research issues and challenges. *IEEE Commun Surv Tuts* 17(1):358–380
- Satyanarayanan M (2017) The emergence of edge computing. *IEEE Comput* 50(1):30–39

Resource Slicing

- [Resource Allocation in SDN/NFV-Enabled 5G Networks](#)

Return Address

- [Mixed Network](#)

RF-Based Energy Harvest

- [Efficient Beamforming Design for Cellular Networks with Energy-Constrained Devices](#)

Robust Games for Power Control in Multi-tier Cellular Networks

Kun Zhu², Ran Wang^{1,2,3}, and Xiangping Zhai²

¹College of Intelligence and Computing, Tianjin University, Tianjin, China

²Nanjing University of Aeronautics and Astronautics, Nanjing, China

³Department of Information and Communication Engineering, Tongji University, Shanghai, China

Introduction

Future wireless networks (e.g., 5G) are expected to be highly heterogeneous composed of macro-cells and a large number of small cells. However, the spectrum-wide cross-tier and co-tier interferences could severely limit the system performance. Accordingly, interference mitigation is of vital importance for which subchannel assignment and power control-based resource allocation have been shown to be a feasible solution for OFDMA-based heterogeneous wireless networks (HetNets).

Centralized resource allocation schemes requiring coordination among all BSs may not be practical in HetNets due to the large amount of small-cell base stations (SBSs) deployed, privacy protection by the owners, and security issues. Accordingly, a noncooperative setting is preferred where each BS is of self-interest and performs resource allocation independently and distributively aiming to maximize its own capacity. Also, a hierarchical structure of the competition among the macrocell base stations (MBSs) and SBSs is considered where the MBSs are assumed to have privileges over SBSs in making decisions, and the mathematical modeling fits naturally into the framework of a Stackelberg game.

Due to the dynamic and random nature of wireless environment and limited amount of coordination in HetNets, information uncer-

tainties naturally appear, and the assumption of perfect information may not be practical. Accordingly, when applying perfect information-based resource allocation schemes in a practical system with parameter uncertainties, possible performance degradation could be experienced. For example, interference constraints are usually imposed to protect high-priority users. However, these interference constraints could be violated due to the uncertainty in channel state information (CSI) in a system using a perfect information-based scheme. Therefore, a robust design of resource allocation is required which takes the imperfect information into account.

To explicitly model the parameter uncertainty, the actual value of the parameter is usually represented by the sum of an estimated nominal value and a perturbation term which is also regarded as the uncertainty part. Similar to that in optimization theory, two approaches are used in game theory literature to deal with the information uncertainties: a Bayesian approach considering average payoff optimization which leads to the development of Bayesian games and a worst-case approach based on robust optimization (Bental and Nemirovski 1998) for worst-case payoff optimization which leads to the development of robust games (Aghassi and Bertsimas 2006).

The Bayesian approach considers the uncertainty term as a random variable the probability distribution of which is assumed to be known, and the performance is guaranteed on an average basis. For the cases where the probability distribution is not available, the distribution-free worst-case approach is a better option which only requires information about the bound of uncertain parameters. Specifically, with the worst-case approach, the uncertainty term is assumed to be bounded within an uncertainty region, and certain level of performance can be guaranteed for every realization in the uncertainty set which can prevent undesirable performance fluctuations.

In this chapter, the uncertainty in CSI of interfering channels is considered following the worst-case approach. Specifically, the downlink power control problem in a two-tier HetNet with

imperfect CSI is formulated as a distribution-free multi-leader multi-follower robust Stackelberg game (RSG), where the MBSs are considered to be the leaders while the SBSs are the followers. The RSG is composed of a leader sub-game and a follower sub-game and accordingly can be expressed as an equilibrium program with equilibrium constraints (EPEC). Robust Stackelberg equilibrium (RSE) composed of a robust Nash equilibrium (RNE) for the follower sub-game, and a RNE for the leader sub-game is considered to be the solution of the RSG (Zhu et al. 2014).

System Model and Game Formulation

System Model

We consider downlink transmissions in an OFDMA-based two-tier cellular network consisting of a set $\mathcal{M} = \{1, 2, \dots, M\}$ of macrocells and a set $\mathcal{K} = \{1, 2, \dots, K\}$ of small cells sharing the same set $\mathcal{N} = \{1, 2, \dots, N\}$ of

orthogonal subchannels. Also, we assume that only one user is served by each MBS/SBS in each subchannel. Denote by $g_{lm}(n)$ the co-tier channel power gain between the MBS l and the user of MBS m on subchannel n . The cross-tier channel power gain between the MBS m and user of SBS k on subchannel n is denoted by $h_{mk}(n)$. Similarly, we can define the co-tier channel power gain between SBSs as $g_{jk}(n)$ and the cross-tier channel power gain between SBS k and user of MBS m as $h_{km}(n)$.

Imperfect CSI is considered which brings uncertainty into the system. Specifically, the actual value of the channel gain is expressed as the sum of an estimated nominal value and an uncertainty term as

$$g_{lm}(n) = \bar{g}_{lm}(n) + \Delta g_{lm}(n), \quad (1)$$

where $\bar{g}_{lm}(n)$ is the nominal value and $\Delta g_{lm}(n)$ is the uncertainty term. In this chapter, ellipsoidal uncertainty model is used due to its analytical tractability. Specifically, the channel uncertainty is described by

$$g_{lm}(n) = \{\bar{g}_{lm}(n) + \Delta g_{lm}(n) : \|\Delta g_{lm}(n)\|_{\mathbf{w}_{lm}(n)} \leq \epsilon_m(n)\},$$

$$= \left\{ \bar{g}_{lm}(n) + \Delta g_{lm}(n) : \sum_{l \neq m} |\Delta g_{lm}(n)|^2 w_{lm}(n) \leq \epsilon_m^2(n) \right\},$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector and $\mathbf{w}_{lm}(n)$ is a vector of positive weights. Similarly, the uncertainty models for $g_{jk}(n)$, $h_{mk}(n)$, and $h_{km}(n)$ can be obtained.

The transmit powers of MBS m and SBS k are denoted by $\mathbf{p}_m = [p_m(1), \dots, p_m(n), \dots, p_m(N)]$ and $\mathbf{p}_k = [p_k(1), \dots, p_k(n), \dots, p_k(N)]$, respectively, where $p_m(n) \geq 0$ and $p_k(n) \geq 0$ are the transmit powers of MBS m and SBS k on subchannel n , respectively. For each MBS and SBS, the total transmit power is limited by

$$\sum_{n=1}^N p_m(n) \leq P_m^{\text{sum}}, \quad \sum_{n=1}^N p_k(n) \leq P_k^{\text{sum}}, \quad (2)$$

where P_m^{sum} and P_k^{sum} are the corresponding power budgets.

The MBSs and SBSs are considered to be of self-interest aiming to maximize their own achievable rate through the allocation of transmit power over the N subchannels subject to certain power and interference constraints. Given the power allocations of all MBSs and SBSs, the interference plus noise experienced by users of MBS m and SBS k on channel n can be expressed as $v_m(n) = \sigma_m^2(n) + \sum_{l \neq m} g_{lm}(n) p_l(n) + \sum_{k=1}^K h_{km}(n) p_k(n)$ and $v_k(n) = \sigma_k^2(n) + \sum_{m=1}^M h_{mk}(n) p_m(n) + \sum_{j \neq k} g_{jk}(n) p_j(n)$, respectively. Accordingly, the maximum downlink transmission rate for MBS m and SBS k can be obtained as follows:

$$\begin{aligned}
\mathcal{R}_m(\mathbf{p}_m, \mathbf{p}_{-m}, \mathbf{p}^{\text{low}}) \\
&= \sum_{n=1}^N \log(1 + p_m(n)/v_m(n)), \\
\mathcal{R}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}}) \\
&= \sum_{n=1}^N \log(1 + p_k(n)/v_k(n)),
\end{aligned}$$

where $\mathbf{p}_{-m}$ and $\mathbf{p}_{-k}$ represent the power allocations of all MBSs except the m th MBS and power allocations of all SBSs except the k th SBS, respectively, $\mathbf{p}^{\text{low}}$ and $\mathbf{p}^{\text{up}}$ represent the vectors of power allocations of all SBSs and MBSs, respectively.

Robust Stackelberg Game Formulation

A robust Stackelberg game is formulated to capture the hierarchical competition among the MBSs and SBSs. Also, the competitive interactions exist within the multiple leaders and the multiple followers. In this case, a robust noncooperative leader sub-game for the leaders and a robust noncooperative follower sub-game for the followers are formulated. Both of them together constitute the robust Stackelberg game. Specifically, the structure of the robust Stackelberg game is described as follows:

- *Players:* The MBSs and SBSs are the players of the robust Stackelberg game. Also, the MBSs are the players (as the leaders) of the leader sub-game, and the SBSs are the players (as the followers) of the follower sub-game.
- *Strategy:* The strategy of each player (both the leaders and the followers) is the selection of power allocation over the N available sub-channels.
- *Strategy space:* Denote by $\mathcal{P}_m$ and $\mathcal{P}_k$ the sets of admissible power allocation of MBS m and SBS k , respectively. The strategy spaces of the leaders and the followers are then given by $\mathcal{P}^{\text{up}} = \prod_{m \in \mathcal{M}} \mathcal{P}_m$ and $\mathcal{P}^{\text{low}} = \prod_{k \in \mathcal{K}} \mathcal{P}_k$, respectively. The strategy space of the entire game is given by the Cartesian product $\mathcal{P}^{\text{up}} \times \mathcal{P}^{\text{low}}$. Note that $\mathcal{P}_m$ and $\mathcal{P}_k$ vary with different considerations of the constraints.

- *Uncertainty models:* Ellipsoidal uncertainty model is considered.
- *Robustness:* For robustness analysis, we consider the worst-case.
- *Payoff:* The payoffs of the MBSs and SBSs are defined as $\pi_m = \mathcal{R}_m(\mathbf{p}_m, \mathbf{p}_{-m}, \mathbf{p}^{\text{low}})$ and $\pi_k = \mathcal{R}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}})$, respectively.

Robust Stackelberg equilibrium (RSE) is considered to be the solution of the game which is constituted by a robust Nash equilibrium (RNE) of the leader sub-game $\mathbf{p}^{\text{up}*}$ and a RNE of the follower sub-game $\mathbf{p}^{\text{low}*}(\mathbf{p}^{\text{up}*})$ and is represented by the tuple $\{\mathbf{p}^{\text{up}*}, \mathbf{p}^{\text{low}*}(\mathbf{p}^{\text{up}*})\}$.

Robust Follower Sub-game

Backward induction is commonly used to analyze a Stackelberg game. Therefore, we start analyzing the robust Stackelberg game by firstly investigating the robust follower sub-game given the strategies of the leaders. Firstly, the admissible strategy set of SBS k is

$$\begin{aligned}
\mathcal{P}_k &= \left\{ \mathbf{p}_k \in \mathbb{R}^N : \sum_{n=1}^N p_k(n) \leq P_k, p_k(n) \geq 0 \right\}, \\
&\forall k \in \mathcal{K}.
\end{aligned}$$

The nominal formulation of the follower sub-game is described as follows. Given the power allocations $\mathbf{p}^{\text{up}}$ of the leaders, each follower k aims to maximize its own achievable downlink transmission rate as

$$\begin{aligned}
&\max_{\mathbf{p}_k} \mathcal{R}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}}) \\
&\text{s.t. } \mathbf{p}_k \in \mathcal{P}_k.
\end{aligned} \tag{3}$$

The robust counterpart of the nominal game can be formulated by considering uncertainties in the channel state information. Specifically, each follower k aims to robustly maximize its downlink rate subject to CSI uncertainty which is a max-min problem described as follows:

$$\max_{\mathbf{p}_k} \min_{\mathbf{g}_k, \mathbf{h}_k} \mathcal{R}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}}) \quad (4)$$

$$\text{s.t. } \mathbf{p}_k \in \mathcal{P}_k, \mathbf{g}_k \in \mathcal{G}_k, \mathbf{h}_k \in \mathcal{H}_k, \quad (5)$$

where $\mathbf{g}_k = \{g_{jk}\}_{j \neq k, j \in \mathcal{K}}$ and $\mathbf{h}_k = \{h_{mk}\}_{m \in \mathcal{M}}$. $\mathcal{G}_k$ and $\mathcal{H}_k$ are the information uncertainty sets

$$\mathcal{G}_k = \left\{ \bar{g}_{jk}(n) + \Delta g_{jk}(n), \|\Delta g_{jk}(n)\|_{\mathbf{w}_{jk}(n)} \leq \epsilon_k^2(n) \right\}_{j \neq k, j \in \mathcal{K}},$$

$$\mathcal{H}_k = \left\{ \bar{h}_{mk}(n) + \Delta h_{mk}(n), \|\Delta h_{mk}(n)\|_{\mathbf{w}_{mk}(n)} \leq \epsilon_k^2(n) \right\}_{m \in \mathcal{M}}.$$

With the ellipsoidal uncertainty model, for worst-case analysis, the problem in (4) can be transformed as

$$\max_{\mathbf{p}_k \in \mathcal{P}_k} \hat{\mathcal{R}}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}}) \quad (6)$$

of player k . Note that since there is only power constraint, the uncertainties appear only in the payoff function.

The ellipsoidal uncertainty sets for follower k are modeled as

$$\hat{\mathcal{R}}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}}) = \sum_{n=1}^N \log \left(1 + \frac{p_k(n)}{\hat{v}_k(n)} \right), \quad (7)$$

and

$$\begin{aligned} \hat{v}_k(n) = & \sigma_k^2(n) + \sum_{m=1}^M \bar{h}_{mk}(n) p_m(n) + \sum_{j \neq k} \bar{g}_{jk}(n) p_j(n) \\ & + \epsilon_k(n) \sqrt{\sum_{j \neq k} \frac{|p_j(n)|^2}{w_{jk}(n)}} + \epsilon_k(n) \sqrt{\sum_{m=1}^M \frac{|p_m(n)|^2}{w_{mk}(n)}}. \end{aligned} \quad (8)$$

The derivation is omitted here for brevity. Given the strategies of the MBSs and all other SBSs, the optimization problem in (6) is still convex. Applying the KKT optimality conditions, a closed-form solution for the robust power control can be obtained, which is a water-filling-like solution, given as

$$\mathbf{p}_k^* = [\mu_k - \hat{v}_k(n)]^+. \quad (9)$$

Robust Leader Sub-game

In the formulated robust Stackelberg game, the leaders need to anticipate the best responses of the followers in terms of the RNE of the follower sub-game. In this case, the leader sub-game becomes a robust equilibrium problem with

equilibrium constraints (EPEC). And for each MBS m , the robust rate maximization problem becomes a mathematical program with equilibrium constraints (MPEC) which can be expressed as follows:

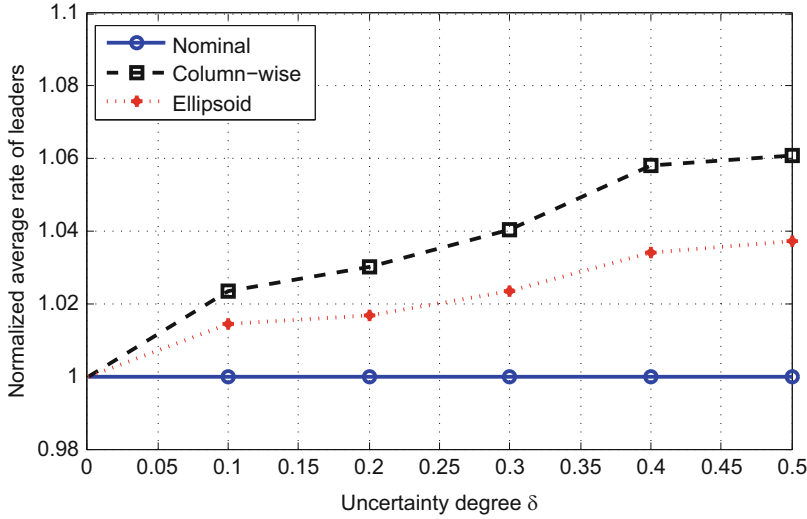
$$\max_{\mathbf{p}_m} \min_{\mathbf{g}_m, \mathbf{h}_m} \mathcal{R}_m(\mathbf{p}_m, \mathbf{p}_{-m}, \mathbf{p}^{\text{low}*}(\mathbf{p}^{\text{up}})) \quad (10)$$

$$\text{s.t. } \mathbf{p}_m \in \mathcal{P}_m, \mathbf{g}_m \in \mathcal{G}_m, \mathbf{h}_m \in \mathcal{H}_m,$$

$$\mathbf{p}_k^*(\mathbf{p}^{\text{up}}) = \arg \max_{\mathbf{p}_k \in \mathcal{P}_k} \min_{\mathbf{g}_k \in \mathcal{G}_k, \mathbf{h}_k \in \mathcal{H}_k}$$

$$\mathcal{R}_k(\mathbf{p}_k, \mathbf{p}_{-k}, \mathbf{p}^{\text{up}}), \forall k, \quad (11)$$

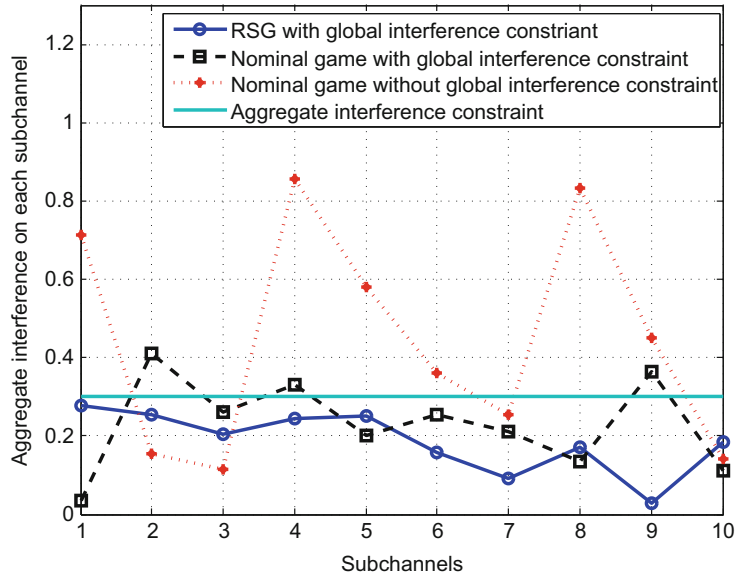
where $\mathcal{P}_m = \left\{ \sum_{n=1}^N p_m(n) \leq P_m, p_m(n) \geq 0 \right\}$ is the power constraint for MBS m and (11) is the equilibrium constraint.



Robust Games for Power Control in Multi-tier Cellular Networks, Fig. 1 The normalized average rate of leaders

Robust Games for Power Control in Multi-tier Cellular Networks, Fig. 2

The aggregate interference experienced by a user of an MBS under different solutions



R

The uncertainty sets for leader m are expressed as

$$\mathcal{G}_m = \left\{ \bar{g}_{lm}(n) + \Delta g_{lm}(n), \|\Delta g_{lm}(n)\|_{\mathbf{w}_{lm}} \leq \epsilon_m^2(n) \right\}_{l \neq m, l \in \mathcal{M}}.$$

The uncertainty sets for the followers are shown in Section III.

Denote by $\mathbf{p}^{\text{up}*}$ an RNE of the leader sub-game, and then the tuple $\{\mathbf{p}^{\text{up}*}, \mathbf{p}^{\text{low}*}(\mathbf{p}^{\text{up}*})\}$ constitutes a robust Stackelberg equilibrium (RSE) which is considered to be the solution of the robust Stackelberg game. We first show the existence of a RSE in the following theorem.

Theorem 1 *There exists at least one robust Stackelberg equilibrium for the formulated*

multiple-leader multiple-follower robust Stackelberg game with ellipsoidal uncertainty model.

Given the leaders' strategies $\mathbf{p}^{\text{up}}$, the followers respond with the unique optimal power allocation $\mathbf{p}^{\text{low}*}$ (assuming that the uniqueness condition of the follower game is satisfied), which can

be numerically obtained. Then each leader m updates its strategy according to the following equation:

$$p_m(n) = \mu_m - \hat{v}_m(n), \quad (12)$$

where

$$\hat{v}_m(n) = \sigma_m^2(n) + \sum_{l \neq m} \bar{g}_{lm} p_l(n) + \sum_{k=1}^K \bar{h}_{km} p_k^*(n) + \epsilon_m(n) \sqrt{\sum_{l \neq m} \frac{|p_l(n)|^2}{w_{lm}(n)}} + \varepsilon_m(n) \sqrt{\sum_{k=1}^K \frac{|p_k^*(n)|^2}{w_{km}(n)}}.$$

Accordingly, an algorithm for reaching the RSE of the game is proposed as shown in Algorithm 1.

Algorithm 1: Algorithm for reaching the RSE

1. Initialization: set $i = 0$, each MBS chooses an initial power allocation $\mathbf{p}_m^i$.
 2. Given the power allocation of all MBSs $\mathbf{p}^{\text{up}}(i)$, each SBS k responds with the unique power allocation $\mathbf{p}_k^*(i)$.
 3. With the best response power allocations of the followers, and given the power allocations of other MBSs, each MBS m updates its power as $p_m^{i+1}(n) = \mu_m - \hat{v}_m^i(n)$.
 4. Repeat steps 2 and 3 until termination condition is satisfied.
-

Numerical Analysis

The performances of robust solutions and nominal solutions are evaluated in a scenario where the CSI is imperfect. The average rates of the leaders are shown in Fig. 1. It can be observed that with uncertainty in channel information, the robust solutions offer better performances than the nominal solutions. The performance gap increases with an increase of the degree of uncertainty.

Figure 2 shows the interference experienced by the leaders in each subchannel. It can be observed that the interference constraint could be violated if nominal solutions are used in an imperfect CSI environment. With the robust

solutions, the interference constraints can be satisfied for each subchannel.

Cross-References

- [Heterogeneous Small Cell Networks](#)
- [5G Wireless](#)

References

- Aghassi M, Bertsimas D (2006) Robust game theory. Math Programm 207(1/2):231C273
- Ben-Tal A, Nemirovski A (1998) Robust convex optimization. Math Oper Res Lett 25(4):769C805
- Zhu, K, Hossain E, Anpalagn A (2014) Downlink power control in two-tier cellular OFDMA networks under uncertainties. IEEE Trans Commun 63(2):520–535

Robustness of Smart Grids

- [Fault Tolerance and Reliability of Smart Grids](#)

Routing

- [Application of Machine Learning in Wireless Sensor Network](#)
- [Delay-Tolerant Network Routing](#)

Routing for Vehicular Ad Hoc Networks

Rui Tian¹ and Baoxian Zhang²

¹Beijing University of Technology, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

Synonyms

Vehicle-to-everything (V2X); Vehicular ad hoc networks

Definition

A self-organized network that uses wireless communication technology to connect mobile vehicles and traffic facilities.

Historical Background

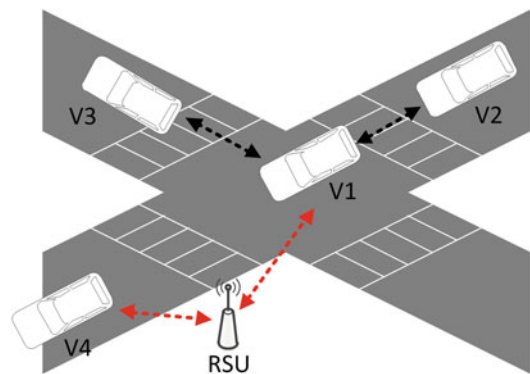
Since early 1980s, wireless communications had been utilized to connect vehicles to support road safety applications. Since there was no wireless communication standard suitable for such scenario, in 1992, dedicated short-range communication (DSRC) was suggested by the American Society to support the wireless communications among vehicles. In 2001, Vehicular Ad Hoc Networks (VANET) was first introduced (Toh 2002) as an application of the mobile ad hoc networks (MANETs). From then on, VANETs had been an independent research field. Latter in 2004, the 802.11p task group in IETF (Internet Engineering Task Force) was formed to draft the 802.11p standard, which became the most significant DSRC technology. By July 2010, after 11 revisions, the IEEE 802.11p standard was published. Later in 2015, a summary of all Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) standards in use was published, including the systems specified by ETSI (European Telecommunications

Standards Institute), IEEE (Institute of Electrical and Electronics Engineers), ARIB (Association of Radio Industries and Businesses), and TTA (Telecommunications Technology Association) (ITU 2015). In 2017, as a contender of 802.11p, LTE-V (Long-Term Evolution-Vehicle) physical layer standards for V2I and V2V communication, which use a radio technology different from IEEE 802.11 WLAN standards, was proposed by the 3GPP.

Foundations

Vehicular ad hoc network, also known as vehicular network utilizing vehicle-to-everything (V2X) communication technology, is a special type of MANET. In MANETs, wireless devices form self-organized wireless networks in movement and transfer data in a multihop manner. More specifically, each device in a MANET can function as packet source, destination, and also router to forward packets for neighboring devices in its communication range. In VANETs, vehicles traveling in cities or along highway act as mobile nodes and form self-organized wireless networks on the fly while moving.

In a VANET, traffic-related and driving safety information can be shared among vehicles in the vicinity directly or with the assistance of fixed road side units (RSUs) (see Fig. 1 for



Routing for Vehicular Ad Hoc Networks, Fig. 1 A typical VANET scenario

illustration purpose). Also, VANETs can work as an extension of the fixed infrastructure to assist vehicles to access the Internet or cellular networks. Massive applications have been deployed or assessed by research institutions and consortia. Typically, VANET applications can be categorized as safety, transport efficiency, and information/entertainment applications (Hartenstein and Laberteaux 2008).

Various wireless technologies can be used in VANETs to support the short-range communication among vehicles and RSUs, such as DSRC, LTE-V2X. Among these standards, DSRC based on IEEE 802.11p has been commercialized by many companies including Toyota, General Motors, and Volkswagen AG (http://www.greencarcongress.com/connected_vehicles/). The IEEE 802.11p standard is an extension of the IEEE 802.11a standard and it was originally designed to handle some special requirements in V2X communication, such as fast message spreading and quick connection establishment in local areas. The Federal Communications Commission (FCC) of USA has allocated 75 MHz bandwidth to V2X communications in the 5.9 GHz frequency band.

Like typical MANETs, in VANETs, data packets are forwarded in a hop-by-hop manner among moving vehicles. Since vehicles move fast and travel in long distance, data sources and destinations may be far away from each other. Since IEEE 802.11p link is range-limited, in most cases, there typically exists no connected routing path between data source and destination (Mohammad et al. 2011). To transfer data in such intermittent connected networks, VANET uses a “store-and-forward” data forwarding pattern which was first introduced in the Inter Planetary Internet (IPN) to transfer data on unreliable links (Fall and Farrell 2008). According to such “store-and-forward” pattern, when there is no available next-hop, current packet holder will temporarily store the packets locally and wait for a communication opportunity to encounter a good next hop. For VANETs, vehicles that act as relay nodes store data packets when moving, so the “store-and-forward” process is also referred to as “store-carry-forward.”

In some cases, the multihop data forwarding in a connected wireless ad hoc network, the forwarding process is also called synchronous routing. In contrast, the “storage-carry-forward” mode in intermittently connected VANET is called asynchronous routing.

For the “store-carry-forward” pattern, the key of routing process is to find next-hop vehicles that are suitable/efficient for delivering data than the vehicle carrying the data packets. For different application scenarios, routing metrics may vary. So it is critical to convert specific routing metrics to routing rules that can assess vehicles’ capabilities of delivering data packets. According to the knowledge required in this process, VANET routing mechanisms can be roughly divided into three categories, namely, replica diffusion based routing, geographical routing, and sociality based routing.

In replica diffusion based routing, vehicles forward data through spreading copies of original data packets in the network. Once a copy is received by the final destination, the message is said to be successfully delivered. In this case, producing more data copies in the network can lead to higher delivery ratio, but also incur excessive resource consumption. In geographical routing, packet holding vehicles always send data to encountered vehicles that are currently driving approaching target nodes’ locations in a hop by hop manner. This type of forwarding strategy requires target nodes to continuously disseminate their fresh position information for relaying vehicles to learn such information during the routing process. For sociality-based routing, since each vehicle’s movement is controlled by a driver, and hence affected by the driver’s social attributes, relaying vehicles are selected from a social view. For example, vehicles that meet each other frequently are also likely to meet each other in future. This is because their drivers are likely to share some common social attributes, such as personal interests, or living and working area. Compared with replica diffusion-based routing and geographical routing, sociality-based routing usually has fewer transmission hops and higher packet delivery rates, but usually experiences much longer delivery delays. For different appli-

cation scenarios, it is necessary to choose one or even combinations of several routing methods to meet actual application requirements.

In addition to general end-to-end data forwarding, there are also more routing demands in VANETs. For example, in a lot of research work, VANET is used as an extension to the fixe infrastructure network to distribute messages on roads, or used as large scale mobile sensing system to collect vehicle status as well as traffic-related information in the city through sensors mounted on vehicles. For data diffusion applications, pub-sub (i.e., publish-subscribe) messaging pattern can be adopted and content-based routing is often preferred in such cases. For large scale data collection applications, vehicles only need to upload their collected data through RSUs distributed in the city, and this type of message routing process is also known as anycast routing, which is different from point-to-point data forwarding, which occurs between specific source-destination pair.

Key Applications

Vehicular Ad Hoc routing technologies are used to support decentralized networking and communication between vehicles, especially for geographical long-distance data forwarding in urban environments.

Cross-References

- [Ad Hoc Networks](#)
- [Delay Tolerant Networks](#)

- [Vehicular Networks](#)
- [Wireless sensor network](#)

References

- Fall K, Farrell S (2008) DTN: an architectural retrospective. *IEEE J Sel Areas Commun* 26(5):828–836
- Hartenstein H, Laberteaux LP (2008) A tutorial survey on vehicular ad hoc networks. *IEEE Commun Mag* 46(6):164–171
- ITU (2015) Recommendation ITU-R M.2084-0: radio interface standards of vehicle-to-vehicle and vehicle-to-infrastructure communications for Intelligent Transport System applications. https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2084-0-201509-!!PDF-E.pdf
- Mohammad S, Rasheed A, Qayyum A (2011) VANET architectures and protocol stacks: a survey. In: *Proceedings of international workshop on communication technologies for vehicles*, Berlin, Heidelberg, vol 6596, pp 95–105
- Toh CK (2002) *Ad hoc mobile wireless networks: protocols and systems*. Prentice Hall, Upper Saddle River

Routing Protocol

- [Energy-Aware and Throughput-Improved Route Selection for Wireless Multi-hop Networks](#)

RSSI-Based Positioning

- [Fingerprinting Localization](#)

Satellite MIMO

Lin Bai¹, Lina Zhu^{2,3}, and Xuejun Zhang¹

¹School of Electronic and Information Engineering, Beijing Laboratory for General Aviation Technology (Beihang University), Beijing, China

²School of Electronic and Information Engineering, Shenyuan Honors College of Beihang University, Beijing, China

³E&CE Department, Illinois Institute of Technology, Chicago, IL, USA

Synonyms

MIMO satellite

Definitions

A promising satellite communication (satcom) technique based on the applications of traditional multiple-input multiple-output (MIMO), which generally includes single-satellite MIMO and multi-satellite MIMO.

Historical Background

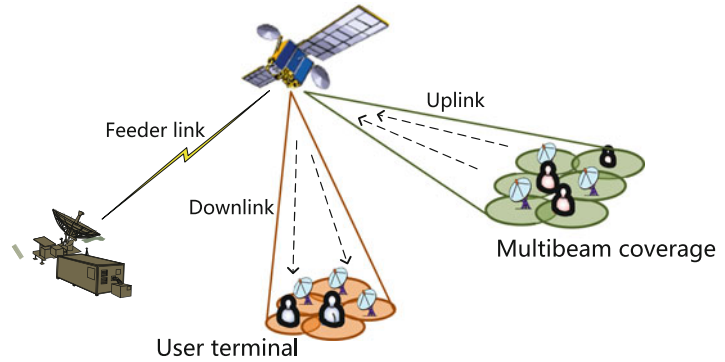
During the past decades, the intensive researches in MIMO technology in terrestrial applications have spanned worldwide with abundant and

spectacular results. Besides time and frequency domain, MIMO-based systems explore and reveal more resources through spatial domain and are able to greatly enhance the system spectrum efficiency via large-scale antennas, without any extra transmission power. It is universally acknowledged that MIMO is one of the most substantial techniques incorporated in the fourth-generation (4G) communication networks to achieve higher throughput and spectrum efficiency, which have been widely standardized in terrestrial communication systems, such as IEEE 802.20, 802.22, 802.11n, 802.16e, 802.16m, etc.

In early studies of satellite architectures, only single antenna with one fixed broad beam is considered. With the increasing demands of communication throughput and services, the great progress in terrestrial MIMO techniques motivates engineers to graft the strengths of multi-antennas onto satcom, which thereby impels the dramatic utilization increase of multibeam satellites to support high-speed transmission with spectrum reuse nowadays (Arapoglou et al. 2010). Moreover, the applications of multi-satellite clustering, as an extension of single-satellite MIMO systems, attract substantial interests of researchers and engineers around the world. Besides the contributions of single-satellite MIMO systems, different satellites can work cooperatively to serve multiple users and form a virtual MIMO system. In this way, the system can benefit from MIMO transmission even with single antenna

Satellite MIMO, Fig. 1

The architecture of single-satellite terrestrial multibeam system



being equipped, which is of great significance for satellites and ground terminals that are not capable for multiple antennas due to hardware or environment restraints.

In future 5G, apart from higher data rate request, it is expected that MIMO performs as flexible support technologies for satellite networks to meet the demands of data distribution, which enable to off-load part of the traffic and optimize the infrastructure dimensioning. Besides, multi-polarization MIMO is suitable for satellite- and context-aware multiuser detection, and massive MIMO plays an important role in satellite multibeam techniques and interference cancelation between multiple adjacent spot beams.

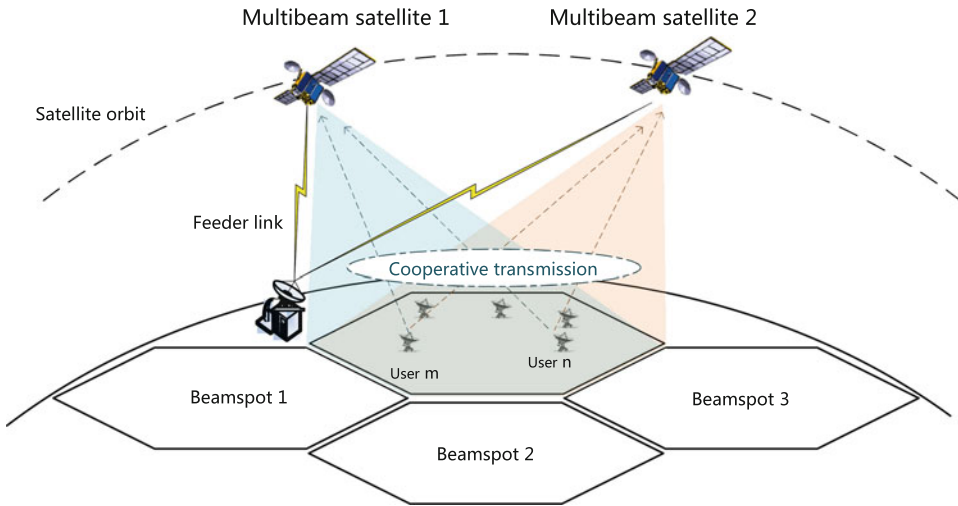
Foundations

Considering the different orbits, number of satellites, or functions of satellite terminals, the MIMO infrastructure between satellites and terrestrial networks should be designed carefully based on service coverage, transmission delay, link geometry, channel impairments, special interference, etc. Besides, the distinct differences of the channel characteristics between satcom and terrestrial communication systems also play significant roles in satellite MIMO applications. For example, the signals in urban areas with rich scatters and obstructions are dominated by multipath components, while in satcom systems, channels are always considered to be the line of sight (LOS) with little scatter.

A typical single-satellite MIMO multibeam system is illustrated in Fig. 1. The satellite generates multiple beams with certain frequency reuse. The ground terminals, assumed to be equipped with only one antenna, are randomly distributed over the area covered by multibeam spots. Users of adjacent beams are allocated with different frequencies. However, thanks to the nature of multibeam, proper frequency reuse pattern can be determined for higher spectrum efficiency.

One of the motivations of multi-satellite transmission is the limited space and spectrum resources in satcom systems, especially in the geostationary orbit. Therefore, letting multiple satellites share the same orbit window greatly releases the burden of orbit and spectrum resources. There are already a lot of studies concentrating on this aspects, such as the works of Robert T. Schwarz since 2008 (Schwarz et al. 2008, 2011). A typical 2×2 multi-satellite MIMO cooperative transmission system where the satellites share the same orbit window is given by Fig. 2. In the system, the selected users transmit signals simultaneously to collocated satellites in the same frequency allocated by the ground control center. Then the satellites forward the received signals to the ground control center through another wide-bandwidth link for signal processing. It is shown that the capacity of satellite MIMO transmission can be further enhanced by using multiple satellites for cooperative transmission.

Another aspect of multi-satellite MIMO transmission is the distributed satellite clusters. This architecture is becoming increasingly attractive in



Satellite MIMO, Fig. 2 The architecture of multi-satellite MIMO cooperative transmission system

the researches of space information networking (Yu et al. 2016), as shown in Fig. 3. The system aims at overcoming communication network heterogeneity and realizing information interaction between different systems in the future. Rather than sharing the same orbit in space, distributed satellite cluster architecture contains a larger map where satellite clusters are separately distributed in different orbits. Satellite clusters can exchange information between each other and work cooperatively to accomplish complex tasks.

Key Techniques

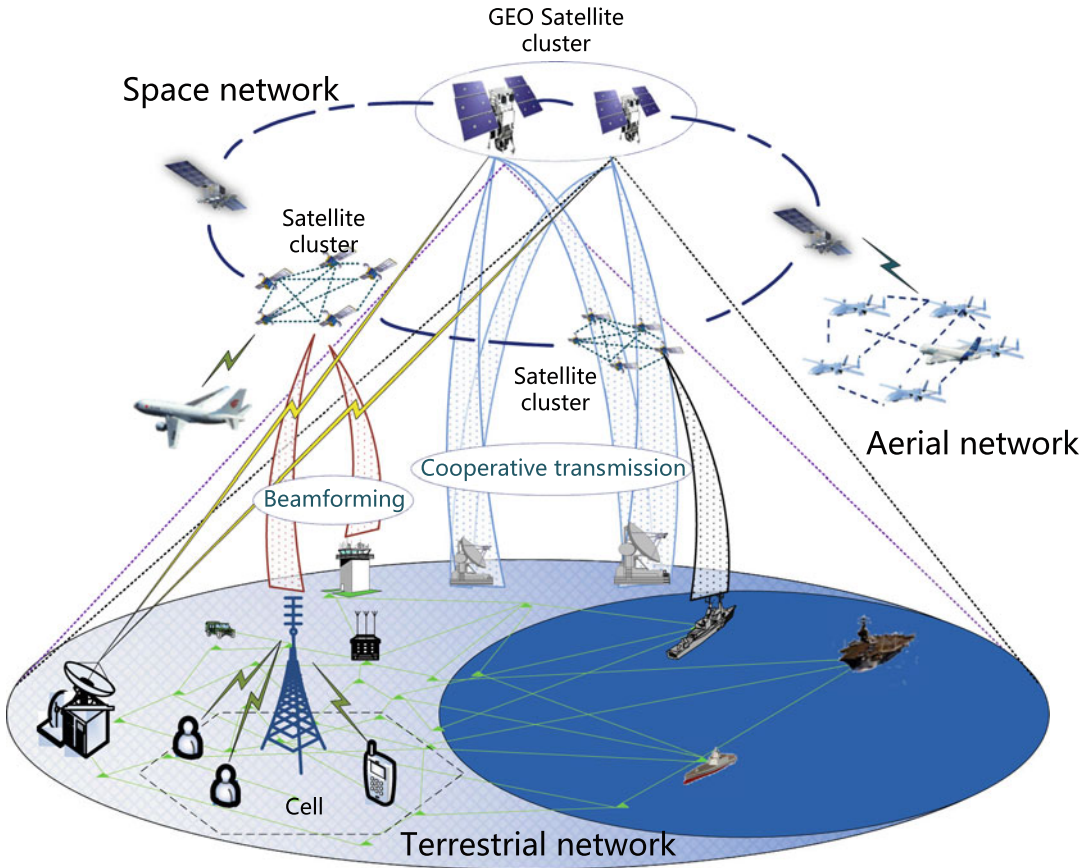
From the architecture of satellite MIMO systems, communication technologies in common terrestrial MIMO systems, such as power allocation strategies, precoding/decoding procedure, user scheduling, etc., are supposed to jointly design to achieve a good balance between resource constraints and the quality of service (QoS) of scheduled user terminals.

As illustrated in Fig. 4, resource allocation and channel estimation are major concerns in single-satellite MIMO transmission. Further, for multi-satellite systems, more advanced schemes such as virtual MIMO and distributed satellite cluster also attract much attention around the world.

Resource Allocation

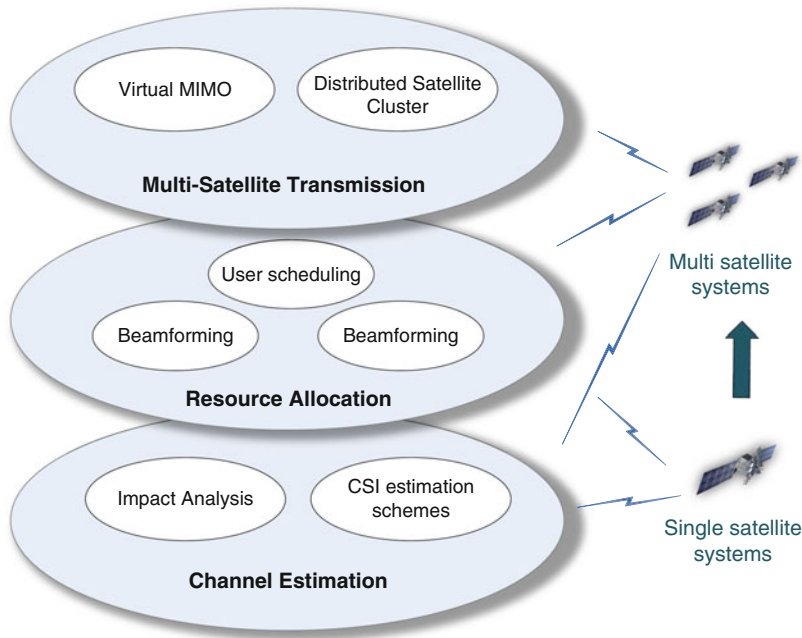
Reasonable resource allocation among multiple users with required QoS is crucial to meet the demand of larger throughput and increasing number of accessing users in future communications. The resource allocation in satcom systems includes power/bandwidth/time-slot allocation, beamforming/precoding schemes, and user clustering/subgrouping.

- Power allocation
 - The evaluating indicators of various power allocation strategies contain the total system throughput, the SINR balancing between different users, and energy efficiency. There have been abundant studies on power allocation in satcom systems, most of which concentrate on one of the indicators, such as sum-rate maximization in certain QoS of users, power minimization, or max-min SINR request. However, the fast evolution of communication markets and the demands of the next-generation targets urge satellite systems to be competitive in the advent of multimedia interactive services. Therefore, the power allocation should be more adaptive and flexible for future satcom systems.



Satellite MIMO, Fig. 3 The architecture of space information networking

- Beamforming/precoding
 - There are two concerns in satellite MIMO beamforming: achieving better capacity performance and mitigating interference (including inter-beam and inter-beam interference), which are usually considered jointly to derive well-designed beamforming/precoding matrixes (Arti and Bhatnagar 2014). For early multibeam satellite systems, beamforming techniques are the extension of well-known terrestrial MIMO beamforming techniques such as linear precoding. With the increasing user scales, subgrouping schemes are used to spatially separate highly correlated users and guarantee orthogonality between different user groups. Afterward, digital beamforming is applied to generate multibeam for different user groups, and other multiple access techniques are used to mitigate intragroup interference.
 - User scheduling
 - A reliable user scheduling strategy should achieve higher transmission quality for each scheduled users, take a good balance between the spectrum efficiency and fairness, and mitigate the inter-beam/inter-cluster interference. In recent researches, more sophisticated scheduling schemes are proposed in satellite multicast systems to realize good balance between performance and complexity, such as the max-min heuristic principle considered in Boussemart et al. (2011).
- Multicast multigroup beamforming is another widely studied field in satcom



Satellite MIMO, Fig. 4 Key techniques in satellite MIMO systems

systems to further reduce interference and improve system performance. Early multicast multigroup beamforming is based on a simple strategy that serves all the users in one spot beam as a single user, and thus the same symbol is broadcasted with multiple users. Afterward, more sophisticated multicast multigroup beamforming schemes based on block singular value decomposition (SVD) (Orsino et al. 2017) or joint optimization, such as frame-based (Christopoulos et al. 2015), are considered to provide better performance.

Channel Estimation

Large amounts of studies demonstrate that the impact of imperfect channel state information (CSI) in satcom systems is the interaction between a series of factors, such as system modeling, channel states (including weather, scattering, etc.), modulation mode, the applied channel estimation strategies, beamforming/decoding schemes, and so on. That results in

degradations of performance such as SER, BER, capacity or effective capacity, etc. (Arti 2016). Meanwhile, multi-antenna structures play a non-negligible role in combat with imperfect CSI. A larger number of antennas can improve the system performance significantly. For satellite MIMO systems, training data-based channel estimation strategies for terrestrial MIMO are still capable, such as symbol-aided (SA) algorithm, symbol-aided plus decision-directed (SADD) algorithm, and decision feedback and adaptive linear prediction (DFALP).

Virtual MIMO and Distributed Satellite Clustering

It has already been demonstrated that orthogonal MIMO channel formulation via the explicit positioning of satellite and user antennas can dramatically improve the capacity of satcom systems, as extra spatial multiplexing gain is obtained by forming virtual MIMO systems between satellites and terminal users. For distributed satellite cluster networks, the resource management between different clusters is a

crucial concern. In comparison with single-satellite systems, the distributed satellite cluster networks can be a dynamic, diverse, and complicated network. Therefore, higher dynamic multi-objective joint allocations of network resource are demanded to satisfy different requirements.

Cross-References

- [Cell-Free Massive MIMO](#)
- [Hybrid Precoding](#)
- [Massive MIMO](#)
- [Massive MIMO Channel Estimation](#)
- [MIMO Relay](#)
- [Network MIMO](#)

References

- Arapoglou PD, Liolis K, Bertinelli M, Panagopoulos A (2010) MIMO over satellite: a review. *IEEE Commun Surv Tutor* 13:27–51
- Arti MK (2016) Channel estimation and detection in satellite communication systems. *IEEE Trans Veh Technol* 65:10173–10179
- Arti MK, Bhatnagar MR (2014) Beamforming and combining in hybrid satellite-terrestrial cooperative systems. *IEEE Commun Lett* 18:483–486
- Boussemart V, Berioli M, Rossetto F (2011) User scheduling for large multi-beam satellite MIMO systems. In: *Conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, Pacific Grove, pp 1800–1804
- Christopoulos D, Chatzinotas S, Ottersten B (2015) Multicast multigroup precoding and user scheduling for frame-based satellite communications. *IEEE Trans Wirel Commun* 14:4695–4707
- Orsino A, Araniti G, Scopelliti P, Gudkova I, Samouylov K, Iera A (2017) Optimal subgroup configuration for multicast services over 5G-satellite systems. In: *IEEE international symposium on broadband multimedia systems and broadcasting (BMSB)*, Cagliari, pp 1–6
- Schwarz RT, Knopp A, Lankl B, Ogermann D, Hofmann CA (2008) Optimum-capacity MIMO satellite broadcast system: conceptual design for LOS channels. In: *Advanced satellite mobile systems*, Bologna, pp 66–71
- Schwarz RT, Knopp A, Lankl B (2011) Performance of an SC-FDE SATCOM system in block-time-invariant orthogonal MIMO channels. In: *IEEE globecom*, Houston, pp 1–6
- Yu Q, Meng W, Yang M, Zheng L, Zhang Z (2016) Virtual multi-beamforming for distributed satellite clusters in space information networks. *IEEE Wirel Commun* 23:95–101

Scalable Smart Grid Communication

- [Smart Grid Communication Based on IEEE 2030 Standard](#)

Scalable Streaming Media

- [Scalable Video Streaming](#)

Scalable Video Coding

- [Scalable Video Streaming](#)

Scalable Video Streaming

Xiaofeng Jiang, Shuangwu Chen, and Jian Yang
University of Science and Technology of China,
Hefei, China

Synonyms

[Scalable streaming media](#); [Scalable video coding](#)

Definition

Scalable video streaming is a new type of adaptive video communication technology based on the scalable video coding and transmission rate control, which make the video communication adaptive to complex multimedia application environments, heterogeneous networks, and different user requirements. The scalable video coding technique is employed to generate the multi-rate streaming videos from the original video files. The transmission rate control technology is employed to dynamically adjust the transmission rate of the streaming videos. Therefore, the

smooth playback can be ensured, while improving the video quality and the user experience.

Historical Background

According to the forecast of American Information Company (Cisco 2014), the video business will account for 80% of the entire Internet data traffic until 2019. From the statistical data, there is an increasing demand for the network multimedia services. Due to the existing problems of the IP-based network architecture, such as the constrained bandwidth and unstable connection, adaptive network management and scheduling strategy is an effective way to design the reliable media service system. The scholars have made many studies, including the adaptive transmission control technology, the adaptive video transcoding, and the media processing technology, such as the scalable video coding (SVC).

The adaptive transmission control and video transcoding allows users to dynamically adjust video frame rates within a certain range. This mechanism can effectively compensate the influence of the network jitter on the video playback, so as to reduce the playback delay, interruption, and improve the user experience. Kalman et al. (2004) and Chuang et al. (2007) used a strategy based on the fixed buffer threshold to trigger the adjustment of the playback rate. Liu et al. (2008) has proposed a H.264 bit allocation strategy based on the rate-distortion (R-D) model. The scalable coding technology encodes videos in the time, space, and quality dimensions as a base layer and multiple enhancement layers. The more video layers the user receives, the higher the corresponding video code rate is, and the better the video quality is. Nejati et al. (2010) established a scalable expected video transmission distortion model in wireless channels and solved the video transmission optimization problem to minimize the distortion.

In recent years, the researches that combine the future network and video streaming has received extensive attentions, which focus on the separation of control and data forwarding, the content centric network, and the in-network

cache. Egilmez et al. (2013) designed a quality of service-enabled (QoS-enabled) scalable video transmission framework for the openflow network. Yang et al. (2018) implemented the adaptive multicast scheme in the software defined network (SDN), where the combination of the media streaming and programmable network can provide better multimedia services. Li and Simon (2011) proposed a collaborative cache strategy for the content centric network (CCN) to alleviate the bandwidth pressure caused by the centralization of a large number of user requests. Rao et al. (2013) considered the active and proactive caching of part data in the named data network (NDN), which was able to better support the user mobility.

Foundations

Architecture of the Network Multimedia Service System

The architecture of the traditional network multimedia service system can be divided into two types, which are the connection-oriented streaming media based on the real time streaming protocol (RTSP) and the connectionless progressive downloading based on the hypertext transfer protocol (HTTP).

In the connection-oriented streaming media system, the server always monitors the client state while sending the media data packet and responds to the client feedback in real time. During the streaming media transmission, the client uses the RTSP protocol to transmit the control commands to the server, such as play, pause, setup, and teardown. The media data packets are carried by the real-time transport protocol (RTP), while the RTP connections are controlled by the real-time transport control protocol (RTCP), and the required network resources are provided by the resource reservation protocol (RSVP). By this way, the online real-time video playback can be guaranteed, and the server can adaptively adjust the media bit rate according to the client feedback. However, the installation, configuration, and maintenance of the streaming media servers are complicated and expensive. HTTP is

a stateless protocol that does not maintain the connection state between the client and the server, and the client can choose to play from any time in the undownloaded video file. The connection-less progressive downloading system is generally applicable to the video on demand services, and is supported by the mainstream terminal players. This framework is scalable, compatible, and flexible to deploy but does not support the live video broadcast services. Due to the lack of media adaptivity, the quality of experience in a wireless network environment is poor.

To solve above problems, the HTTP adaptive streaming (HAS) technology has emerged. Since the media data is transmitted by the HTTP, it has the good compatibility, strong scalability, and flexible deployment. Meanwhile, the same media content is encoded into multiple code rate counterparts, the client can select the appropriate version based on its own network conditions and improve the media adaptivity.

Media Adaptive Technology

Recently, the high-definition multimedia services are required to provide the higher quality of service (QoS) and the better quality of experience (QoE). However, different terminals have different resolution and decoding capabilities, and different access networks have different bandwidths, delays, and packet loss rates, which pose a huge challenge to the development of network multimedia services. To overcome this challenge, the adaptivity of the media content is becoming more and more important. The following three methods can be used to improve the adaptivity.

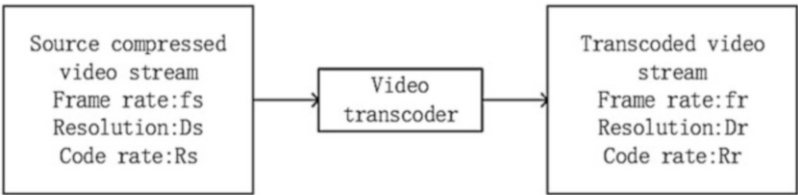
The online adaptive transcoding method is a technology that dynamically changes the encoding parameters, such as the frame rate, spatial resolution, and the like, before transmitting the

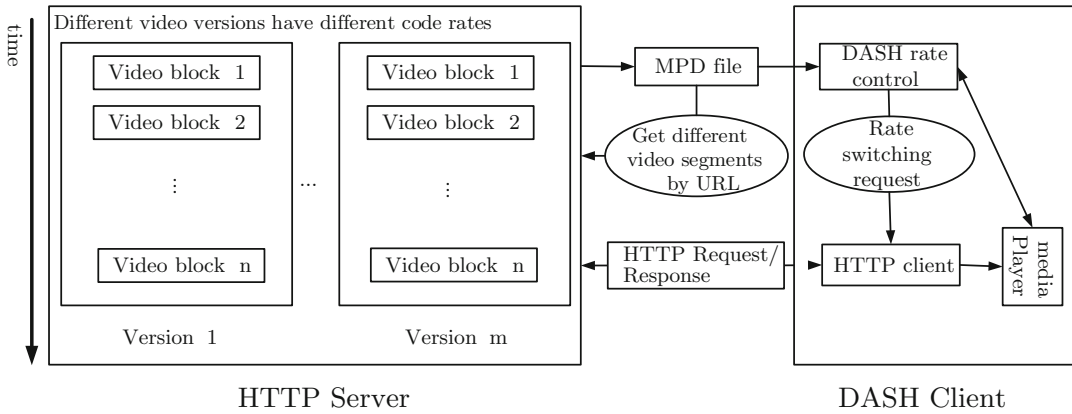
compressed video stream to the client. The original stream is thus converted into a different video stream with different frame rates and resolutions. The schematic diagram is shown in Fig. 1. By adjusting the frame rate, spatial resolution, and image quality, the method can convert the high quality video stream into a low-quality version, or convert the low-quality video stream into a high-quality version. Therefore, the online transcoding can enhance the media adaptivity, support heterogeneous networks and terminals, different systems and formats.

The SVC is a method that can layer code the video stream in space, time, and quality, which has been included in the extended version of the H.264/AVC video codec standard and is known as the H.264-AVC standard. With the SVC, the original video is encoded into a multi-layer video stream with one base layer and multiple enhancement layers. The base layer has a small amount of data, low bandwidth requirements, and can be independently decoded to provide a barely acceptable viewing quality. As the number of enhancement layers increases, the bandwidth requirements increase gradually. Meanwhile, depending on the layering method, the video quality can be improved from the resolution, frame rate, and image quality. This layered approach greatly enhances the media adaptivity. The SVC video stream can support multiple heterogeneous networks and terminals, and can dynamically adjust the video stream as the network bandwidth changes.

The HAS system uses a multi-rate block coding method to enable media adaptation. In Fig. 2, a dynamic adaptive streaming over HTTP (MPEG-DASH) architecture is used to show the method. After the original video is encoded by the DASH encoder, it is compressed into a plurality of counterparts, which have the same

Scalable Video Streaming, Fig. 1 Online transcoding process





Scalable Video Streaming, Fig. 2 General structure of MPEG-DASH

video content but different code rates and quality. Each version of the video stream is sliced into a number of small video blocks in the time domain, and each video block is a file that can be decoded independently. The original video generates the corresponding media presentation description (MPD) file after encoding. Each video block corresponds to a uniform resource locator (URL) and is recorded in the MPD. When starting to transmit the video content, the client senses the network bandwidth to dynamically select an acceptable video bit rate, which is determined by comparing the bandwidth requirements of the different video versions in the MPD file. Then the serial number of the next video block after the current playing time is determined, and the URL of the corresponding video block can be found in the MPD file. Finally, the client can use the URL to initiate a video request. Since the video blocks are sequentially requested by the client in the time domain, the content between the video blocks is continuous and does not overlap, the video is seamless and smooth for the user to play.

QoS Evaluation of Adaptive Media

Traditional QoS metrics include the network latency, throughput, packet loss rate, bit error rate, jitter, and etc. These indicators reflect the performance of the network transmission layer but ignore the video characteristics and the subjective feelings of the users. Some QoS indicators from the user's perspective should be

defined to fully describe the adaptive multimedia services.

Ignoring the impact of users and the environment, when the video sources used are the same, the QoS mainly depends on factors, such as the video quality, smoothness of playback, and video. The video quality can be calculated by comparing each pixel of the original video frame with the distorted video frame to get the similarity or fidelity, which is represented by the average mean square error (MSE) and peak signal to noise ratio (PSNR),

$$MSE = \frac{\sum_{0 \leq i \leq M} \sum_{0 \leq j \leq N} (f_{ij} - f'_{ij})^2}{M \times N}, \quad (1)$$

$$PSNR = 10 \lg \frac{(2^n - 1)^2}{MSE}, \quad (2)$$

where M , N represent the height and width of the video frame, f_{ij} and f'_{ij} represent the original video frame and the distorted video frame, and n is the number of bits per sample. The video fluency also seriously affects the user's willingness to watch. The video fluency is mainly reflected by the interruption and the stuck in the playback process. Therefore, the average frame rate f_T and the interruption frequency f_C can be used to characterize the fluency in the practical applications,

$$f_{T'} = \frac{S(T)}{T'}, f_C = \frac{C(T)}{T}, \quad (3)$$

where $S(T)$ indicates the total number of frames when playing a video of length T , T' is the actual playback time, and $C(T)$ denotes the total number of interruptions during the playback process. The video viewing experience of the user depends not only on the image quality per frame but also on the quality fluctuation. Rapid image quality fluctuations will lead to the video flicker and reduce the QoE. For the smoothness of the video, the variance of the quality or code rate can be used to characterize,

$$\sigma^2 = \frac{\sum_{t=1}^T (r_t - \bar{r})^2}{T} \quad (4)$$

where T is the video duration, r_t is the code rate at time t , and $\bar{r}$ is the average code rate.

In addition to the QoS evaluation methods listed above, evaluating QoE from a user perspective is also an effective approach. Recently, the structural similarity (SSIM) based on a single image can quantize the video QoE, that is, by mapping the time-averaged SSIM onto the mean opinion score (MOS), the QoE can be obtained.

Key Applications

The scalable video streaming can be applied in the following two aspects:

- *Adaptive transmission of scalable video in wireless networks*

In a wireless network environment, the user mobility, channel interference, and time-varying background traffic will cause the random bandwidth changes. Therefore, the media rate adaptive algorithm based on the underflow probability is necessary for video transmission in a wireless network. The algorithm takes into account the bandwidth changes, the buffer state, the flicker effect on the QoE, and does not rely on the prior knowledge of any wireless channel and video information. Due to the discrete nature of

the scalable video bit rate, blindly increasing the number of layers is easy to cause the oscillation of the video quality. In order to overcome the oscillation, the variation of the code rate is defined as the perturbation factor, and the perturbation analysis model is introduced to evaluate whether the change of the code rate causes the oscillation of the quality. Thereby the blind switching of the video can be avoided and the switching rate can be effectively reduced. The adaptive algorithm based on the underflow probability can significantly improve the smoothness of the video playback while ensuring the video quality.

- *Joint optimization of routing and rate control under SDN architecture*

The SDN architecture has the characteristics that the network state is observable and the route can be dynamically configured. The joint video rate control and dynamic routing optimization algorithm provides self-aware, self-configuring, and self-optimizing multimedia services under the SDN architecture. The algorithm makes decisions based on the currently observed network state, such as the topology, throughput, and packet loss rate, without any prior information. In the video transmission process, the algorithm can adaptively adjust the bit rate and routing path based on the network conditions, which significantly enhances the QoS. Meanwhile, with the multipath transmission mode, the idle bandwidth resources in different paths can be effectively utilized to improve the video quality. Comparing with the fixed rate and routing algorithms, the joint optimization algorithm can effectively improve the video quality, reduce the packet loss rate, and increase the utilization of the bandwidth resources.

Future Directions

The emergence of various new technologies makes the network management and configuration more flexible and also provides a new opportunity for the development of network multimedia technology. Combining traditional

streaming technology with new network technology can optimize the transmission quality of multimedia services and improve the utilization efficiency of network resources, so as to establish self-aware, self-configuring, and self-optimizing network multimedia service system. The future work mainly focus on the following aspects:

- Multimedia service based on intra-network caching

With new network technologies, the routers have the computing and storage capabilities to process and cache all or part of the passing data. During the routing process, the cache on the router can be directly used as a data source to respond to the service requests, thereby avoiding the network congestion from the original data source, shortening the acquisition delay of the media content, and reducing the repeated transmission of data packets. Therefore, an appropriate caching strategy can enhance the resource sharing, improve the network utilization, and speed up the distribution of multimedia content. The core problem is to study the selection strategy of the caching nodes, the size of the allocated cache, and the media content to cache on the basis of the requests from the massive users. In addition, considering the mobility of users, proactively push the media content to the user's neighboring access nodes to solve the problem of seamless switching is also need to be concerned.

- Content-aware adaptive multimedia service

By naming the media content and separating the content identifier from the network address, the content attributes can be recognized, such as the real-time requirements, popularity, preset priorities, and etc. By introducing the measurement, feedback, prediction techniques, and stochastic optimization methods in modern control theory, the dynamic perception and prediction of the network resources and traffic load can be performed. Combining the network and media content attributes, the appropriate caching and routing strategy can be developed

to allocate the network resources, adjust the video bit rate, and optimize the transmission path.

Cross-References

- [HTTP Adaptive Streaming](#)

References

- Chuang H, Huang C, Chiang T (2007) Content-aware adaptive media playout controls for wireless video streaming. *IEEE Trans Multimedia* 9(6):1273–1283
- Cisco I (2014) Cisco visual networking index: forecast and methodology, 2014–2019. CISCO White paper 518
- Egilmez H, Hilmi E, Civanlar S, Tekalp A (2013) An optimization framework for qos-enabled adaptive video streaming over openflow networks. *IEEE Trans Multimedia* 15(3):710–715
- Kalman M, Steinbach E, Girod B (2004) Adaptive media playout for low-delay video streaming over error-prone channels. *IEEE Trans Circuits Syst Video Technol* 14(6):841–851
- Li Z, Simon G (2011) Time-shifted tv in content centric networks: the case for cooperative in-network caching. In: *Proceedings of 2011 IEEE international conference on communications (ICC)*, 5–9 June 2011, Kyoto, pp 1–6
- Liu Y, Li Z, YC S (2008) Rate control of h.264/avc scalable extension. *IEEE Trans Circuits Syst Video Technol* 18(1):116–121
- Nejati N, Yousefi-zadeh H, Jafarkhani H (2010) Distortion optimal transmission of multi-layered fgs video over wireless channels. *IEEE J Sel Areas Commun* 28(3):510–519
- Rao Y, Zhou H, Gao D, Luo H, Liu Y (2013) Proactive caching for enhancing user-side mobility support in named data networking. In: *Proceedings of 2013 seventh international conference on innovative mobile and internet services in ubiquitous computing*, 3–5 July 2013, Taichung, pp 37–42
- Yang J, Yang E, Ran Y, Bi Y, Wang J (2018) Controllable multicast for adaptive scalable video streaming in software-defined networks. *IEEE Trans Multimedia* 20(5):1260–1274

Scheduling

- [Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks](#)

SCTP

- [Handoff Management in Wireless Communication-Based Train Control Systems](#)

SDN-Enabled Cloud Data Center

- [SDN-Enabled Data Center Networks](#)

SDN-Enabled Data Center Networks

Rongtao Xu¹ and Zhe Yang²

¹State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

²Intelligent Computing and Communications (IC²) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Synonyms

[SDN-enabled cloud data center](#); [Software-defined data center](#)

Definitions

Software-defined network (SDN) is an emerging networking paradigm which separates the network's control logic (the control plane) from the underlying routers and switches (the data plane) and enables the network to be programmable, adjustable, and centralized controlled.

Historical Background

The history of SDN can be traced back to the active networks, which were widely studied

from the mid-1990s to the early 2000s. These techniques introduced programmable functions to dynamically adjust network configurations (Feamster et al. 2013). From around 2001 to 2007, the concept of control and data plane separation was proposed. In 2008, McKeown developed a standard application programming interface (API) between the control and data planes, i.e., the OpenFlow, which was regarded as the first practical and scalable instance of widespread SDN open interface (McKeown et al. 2008). Additionally, the Open Networking Foundation was founded in 2011 to further promote the development of OpenFlow and SDN. Recently, due to the high scalability and adjustability, SDN is adopted to data center networks (DCNs) to promote traffic transfer and management among servers (Son et al. 2018; Hafeez et al. 2017).

Foundation

A typical SDN consists of three main components: the centralized SDN controller, SDN-enabled switches, and hosts (Cui et al. 2016). The controller manages the control plane and is responsible for installing flow rules in the data planes of the network devices. It can also obtain the real-time states of the network using probes for specific network statistics. Generally, SDN architecture includes two main APIs, i.e., the southbound API and the northbound API. The former defines the interface between a controller and switches, while the latter defines the interface exposed by the controller to the network applications. DCN usually includes many switches and servers, which may generate large numbers of data flows. Therefore, some challenges may restrict the performance of DCN, e.g., the traffic congestion, energy consumption, and malicious attacks. SDN technology has been regarded an effective method to deal with the above challenges. Existing studies of SDN-enabled DCNs mainly consider three objectives, i.e., congestion control, energy efficiency, and security.

Congestion Control

Based on the state of the entire network, the centralized controller of SDN is able to provide fine-grained control in order to mitigate the traffic congestions in DCN. For example, the controller can detect the occurrence of traffic congestion and direct the traffic flows to a vacant path.

In Gholamiand et al. (2015), the SDN controller firstly gets statistics of ports and queues regularly. Once the buffer occupancy of a certain switch exceeds a preset threshold, the rerouting is conducted. The authors utilized OpenFlow platform to validate the proposed scheme. Moreover, as flows can be categorized into elephant flows (large size) and mice flows (small size), congestions control schemes on them are also different (Hafeez et al. 2017). For instance, Kanagevlu proposed a rerouting approach which identifies and labels flows larger than 100 KB (Kanagevlu et al. 2015). If the load of a certain link exceeds a preset threshold, the controller could install new flow rules to reroute several elephant flows to an alternate path.

Energy Efficiency

As DCN consumes enormous amounts of electricity, methods to improve energy efficiency have received much attention recently. Thanks to the SDN, the network elasticity is easy to be achieved.

In Heller et al. (2010), the authors proposed a power-saving scheme for the fat-tree network topology. Using this scheme, data flows can be consolidated into a smaller number of links, which provides possibilities to turn off unused switches and thus save considerable amount of energy. Moreover, the authors also evaluated the relationship among energy efficiency, transmission efficiency, and network robustness, under various traffic patterns extracted from real-world data traces. X. Wang et al. also studied the traffic consolidation-based energy-saving scheme for DCN (Wang et al. 2016). In this scheme, the correlation analysis is conducted based on the variation trends of traffic flows. Using the analysis results, the controller is able to place the flows with similar trends (i.e., hit the peak or trough at the same time) on the same

part of the network and dynamically turn off unused switches. Aujla et al. presented an SDN-based energy-aware flow scheduling algorithm, which considered the energy efficiency as one of the optimization objectives (Aujla et al. 2018). The multi-leader and multi-follower Stackelberg game theory was used to model and solve the problem. Evaluations were conducted on the Google workload traces to show the effectiveness of the proposed algorithm.

Security

DCN is also vulnerable for some malicious attacks, e.g., the distributed denial of service (DDoS) attack. By utilizing specific SDN functions, e.g., traffic analysis, centralized control, and dynamic update of network rules, the DCN can provide fast detection and reaction to malicious attacks.

In Xu et al. (2016), the authors designed an SDN-based DDoS detection system for data centers. The system consists of two parts: victim detection and post-detection. Firstly, the IP addresses of both victims and attackers are detected based on two flow features, i.e., flow volume and flow rate asymmetry. Then, rules can be installed in the SDN-enabled switches in order to drop packets from attackers to victims. In this way, the effects of DDoS attacks are significantly mitigated. Moreover, Chowdhary et al. proposed a security framework to prevent DDoS attacks for SDN-enabled DCN (Chowdhary et al. 2017). Due to the programmability of SDN, the authors designed a reward and punishment scheme using game theory. This scheme made the bandwidths of potential attackers to be downgraded for a certain period of time and restored back to normal if they resumed cooperation. Evaluations were conducted on various network topologies, e.g., linear topology and fat-tree topology.

Key Applications

SDN-enabled data center networks can provide high-performance, energy-efficient, and security solutions for cloud computing and big data applications.

Cross-References

- [Data Center Network Virtualization](#)
- [Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks](#)

References

- Aujla G et al (2018) Optimal decision making for big data processing at edge-cloud environment: an SDN perspective. *IEEE Trans Ind Inf* 14(2):778–789
- Chowdhary A et al (2017) Dynamic game based security framework in SDN-enabled cloud networking environments. In: *Proceedings of the ACM SDN-NFV*, pp 53–58
- Cui L et al (2016) When big data meets software-defined networking: SDN for big data and big data for SDN. *IEEE Netw* 30(1):58–65
- Feamster N et al (2013) The road to SDN: an intellectual history of programmable networks. *ACM Netw* 11(12):1–13
- Gholamiand M et al (2015) Congestion control in software defined data center networks through flow rerouting. In: *Proceedings of the 23rd Iranian conference on electrical engineering (ICEE)*, pp 654–657
- Hafeez T et al (2017) Detection and mitigation of congestion in SDN enabled data center networks: a survey. *IEEE Access* 6:1730–1740
- Heller B et al (2010) ElasticTree: saving energy in data center networks. In: *Proceedings of 7th USENIX conference on networked systems design and implementation (NSDI10)*, pp 17–22
- Kanagevlu R et al (2015) SDN controlled local re-routing to reduce congestion in cloud data center. In: *Proceedings of international conference on cloud computing research and innovation (ICCCRI)*, pp 80–88
- McKeown N et al (2008) OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM Comput Commun Rev* 11(2):69–74
- Son J et al (2018) A taxonomy of software-defined networking (SDN)-enabled cloud computing. *ACM Comput Surv* 51(3):1–36
- Wang X et al (2016) Correlation-aware traffic consolidation for power optimization of data center networks. *IEEE Trans Parallel Distrib Syst* 27(4):992–1006
- Xu Y et al (2016) DDoS attack detection under SDN context. In: *Proceedings of IEEE INFOCOM*, pp 1–9

Second-Order Reaction

- [Reaction-Diffusion Channels](#)

Secret Sharing

- [Wireless Group Authentication](#)

Secure

- [Smart Grid Communication Based on IEEE 2030 Standard](#)

Secure Association

- [Secure Device Pairing](#)

Secure Data Collection

- [Privacy-Preserving Data Collection](#)

Secure Device Pairing

Hsu-Chun Hsiao
Department of Computer Science and
Information Engineering, National Taiwan
University, Taipei, Taiwan

Synonyms

[Secure association](#)

Definition

Device pairing (*pairing* for short) is the process of establishing a security association between two devices that share no prior knowledge. The security association, such as a shared secret key, is known only to the two devices and thus can

be used to protect subsequent communication between them.

Secure device pairing requires the pairing process to be resilient against a *man-in-the-middle* (MitM) attack, where an adversary attempts to impersonate at least one of the devices by eavesdropping on, intercepting, manipulating, or relaying messages between them.

Secure device pairing and *secure association* are frequently used interchangeably. Since pairing typically involves only two devices, *group device pairing* or *group association* has been used to indicate establishing a group key among three or more devices.

Background

Secure device pairing is crucial for initiating secure communication between previously unassociated devices. After performing secure device pairing, both devices can be assured of an established shared secret with the intended party, even in the presence of *man-in-the-middle* (MitM) attacks. Such an MitM adversary can overhear, intercept, manipulate, or relay messages sent between the devices.

Secure device pairing over a wireless channel with no infrastructure support is more challenging than over a wired one. This is due to the broadcast nature of wireless mediums, which makes the wireless channel highly susceptible to eavesdropping and manipulation. Furthermore, when applied to resource-constrained devices (e.g., a device with no output display), or devices operated by ordinary users without technical expertise, the pairing method must minimize additional input/output hardware and avoid complex user interaction. Hence, a long-standing challenge for secure pairing methods in wireless ad hoc settings is to achieve security, usability, and deployability simultaneously. For comparative surveys of secure device pairing methods, see Kainda et al. (2009), Kobsa et al. (2009), Kumar et al. (2009), Mirzadeh et al. (2014), Chong et al. (2014), and Fomichev et al. (2018).

Foundations

Secure device pairing methods can be broadly divided into two types, by whether a secondary channel is present or not. The secondary channel is typically referred to as an *out-of-band* (OOB) channel, opposite to the primary *in-band* channel. While the in-band channel is vulnerable to *man-in-the-middle* (MitM) attacks, the OOB channel is *assumed* to guarantee authenticity despite MitM threats. For example, the historically important secure device pairing protocols that were proposed in the 1990s and 2000s, including Resurrecting Duckling (Stajano and Anderson 1999), Talking to Strangers (Balfanz et al. 2002), Seeing-is-Believing (McCune et al. 2005), and Loud-and-Clear (Goodrich et al. 2006), use physical contact, infrared, visual, and audio channels as the OOB, respectively. Because the OOB channel usually has limited bandwidth capacity, it is mainly used for authenticating a small amount of data, such as the cryptographic hash of a public key, but is not meant to replace the high-capacity in-band channel.

Secure device pairing with OOB. After obtaining authentic information via OOB, the devices can then perform *authenticated key exchange* via the in-band channel to establish a shared secret. To understand how OOB facilitates secure device pairing in general, let us consider a common scenario in which the devices would like to establish a shared secret by performing the *Diffie-Hellman* (DH) key exchange over an in-band channel. The DH protocol is widely adopted and is resilient against passive eavesdroppers overhearing exchanged messages. However, the DH protocol is known to be vulnerable to MitM attacks, in which the attacker actively interacts with both devices and can intercept, manipulate, or relay messages in addition to eavesdropping. More specifically, without using an OOB channel to obtain authentic DH public keys, a MitM attacker can easily impersonate both devices without being detected, and the attacker establishes shared keys with each device separately by relaying messages between them both. On the other hand, by exchanging the DH public keys via an OOB channel, both devices can verify

whether they obtain the correct DH public key of the intended party. If the OOB channel not only guarantees authenticity but also secrecy, the devices may also derive a shared secret directly from OOB.

A wide spectrum of OOB channels have been explored, including Wi-Fi, Bluetooth, mm-waves, radio-frequency identification (RFID), near-field communication (NFC), visible light communication (VLC), visual, Infrared Data Association (IrDA), audio, ultrasound, haptic, and sensing (Fomichev et al. 2018). As explained above, OOB channels are *assumed* to guarantee authenticity despite MitM threats. When utilizing an OOB channel, a secure device pairing method must justify this assumption under reasonable application scenarios and adversary models. For example, NFC might be applicable as an OOB channel if the attacker is outside NFC's communication range and the devices are equipped with corresponding transceiver modules. As another example, the Line-of-Sight property provided by VLC and infrared makes it easy to detect MitM attacks, assuming that a user is present and vigilant enough to spot potential MitM attackers physically standing between the two devices.

Many secure pairing methods rely on human-perceivable OOB, including visual, audio, or haptic channels. The user must actively verify whether both devices agree on the exchanged information by comparing the displayed codes or lighting patterns on both devices, typing the displayed PIN on one device into the other, etc. For example, the pairing protocol in the Bluetooth Low Energy (BLE) standard supports two options that require active user involvement. In the *Passkey Entry* option, a 6-digit PIN code is displayed on one device, and the user inputs the PIN to the other. In the *Numeric Comparison* option supported by BLE4.2+, both devices display a 6-digit PIN code, and the user checks if the PIN codes match.

Because these pairing methods must interact with the user, they may be inapplicable for devices with limited input/output capabilities (e.g., ones without screens to display codes). In addition, the security of these pairing methods depends on the assumption that the user can

correctly follow the instructions, but in reality procedures involving users are often error-prone. As a result, when utilizing human-perceivable OOB, security is sometimes sacrificed in exchange for better user experience, such as a shorter pairing delay.

To reduce or even avoid user involvement, researchers have proposed secure pairing methods that rely on physically constrained OOB channels, which are assumed to be resilient against MitM attacks because of inherent physical constraints. For example, near-field communication (NFC) is a short-range radio-frequency channel typically operating in centimeter-level distance and thus can act as a location-limited OOB resilient against a remote MitM attacker.

Some physically constrained OOB channels leverage the similar contexts of co-present devices for authentication and even key establishment. The similarity in contexts can be seen as proof of physical proximity and can be used for deriving a shared secret key. For example, the user may be asked to shake both devices together, such that their accelerometer readings can be used to authenticate the devices and create a shared secret. As another example, the user can physically hold both devices at the same time, and a secret can be established via intra-body communication. Some context-based pairing methods require no user interaction by harvesting ambient contexts (Miettinen et al. 2014), such as acoustic, electromagnetic, luminosity, or channel state information (CSI), to name a few. As ambient sensing may take a long time to achieve a sufficient level of robustness and security, pairing methods often combine multiple sensing channels.

In-band only (secure device pairing without OOB). Because requiring a secondary channel might be impractical in some application scenarios, researchers have also considered in-band-only solutions to secure pairing (Gollakota et al. 2011). To mitigate MitM adversaries, in-band-only approaches often rely on the channel's physical properties and propose specific encoding or modulation schemes such that any alteration can be easily detected.

Secure pairing in a group. Several proposals extend the secure pairing problem to group settings (Lin et al. 2009). One variant considers secure pairing in a group, where devices want to obtain each other's authentic information, such as public keys. Another variant considers establishing one shared secret among all devices belonging to the same group. Although secure pairing in a group of n devices can be straightforwardly reduced to performing n^2 pairwise pairing, or n pairwise pairing with a trusted leader, they suffer from high latency and user burden. In addition, advanced attacks, such as group partition, need to be taken into consideration.

Device pairing in standardized wireless protocols. Bluetooth Low Energy (BLE) supports four authentication methods to secure key exchange during device pairing. Devices can negotiate and select one to use based on their I/O capabilities.

- *Just Works* provides no authentication.
- *Passkey Entry* displays a 6-digit PIN code on one device, and the user needs to type the PIN on the other.
- *Numeric Comparison*, introduced in BLE version 4.2, displays 6-digit PIN codes on both devices, and the user needs to confirm that they are identical.
- *Out-of-band (OOB)* is directly provided with a shared key and assumes the key has been securely established via an alternative OOB channel.

Wi-Fi Protected Setup supports four modes for pairing a new wireless device and access point: *PIN*, *Push Button*, *NFC*, and *USB*. While supporting the PIN mode is mandatory, it has been shown that PIN-based WPS is vulnerable to simple brute force attacks due to flawed protocol design and should be disabled for security.

Key Applications

Secure device pairing is mainly used for bootstrapping secure communication between previ-

ously unassociated devices in an ad hoc manner.

Cross-References

- [Secure Association](#)

References

- Balfanz D, Smetters DK, Stewart P, Wong HC (2002) Talking to strangers: authentication in ad-hoc wireless networks. In: Network and distributed system security symposium (NDSS)
- Chong MK, Mayrhofer R, Gellersen H (2014) A survey of user interaction for spontaneous device association. *ACM Comput Surv (CSUR)* 47:8:1–8:40
- Fomichev M, Alvarez F, Steinmetzer D, Gardner-stephen P, Hollick M (2018) Survey and systematization of secure device pairing. *IEEE Commun Surv Tutor* 20(1):517–550
- Gollakota S, Ahmed N, Zeldovich N, Katabi D (2011) Secure in-band wireless pairing. In: USENIX security
- Goodrich MT, Sirivianos M, Solis J, Tsudik G, Uzun E (2006) Loud and clear: human-verifiable authentication based on audio. In: IEEE international conference on distributed computing systems (ICDCS)
- Kainda R, Flechais I, Roscoe A (2009) Usability and security of out-of-band channels in secure device pairing protocols. In: ACM symposium on usable privacy and security
- Kobsa A, Sonawalla R, Tsudik G, Uzun E, Wang Y (2009) Serial hook-ups: a comparative usability study of secure device pairing methods. In: ACM symposium on usable privacy and security
- Kumar A, Saxena N, Tsudik G, Uzun E (2009) A comparative study of secure device pairing methods. *Pervasive Mob Comput* 5:734–749
- Lin Y-H, Studer A, Hsiao H-C, McCune J, Wang K-H, Krohn M, Lin P-L, Perrig A, Sun H-M, Yang B-Y (2009) SPATE: small-group PKI-less authenticated trust establishment. In: ACM international conference on mobile systems, applications, and services (MobiSys)
- McCune J, Perrig A, Reiter M (2005) Seeing-is-believing: using camera phones for human-verifiable authentication. In: IEEE symposium on security and privacy
- Miettinen M, Asokan N, Nguyen TD, Sadeghi A-R, Sobhani M (2014) Context-based zero-interaction pairing and key evolution for advanced personal devices. In: ACM computer and communications security (CCS)
- Mirzadeh S, Cruickshank H, Tafazolli R (2014) Secure device pairing: a survey. *IEEE Commun Surv Tutor* 16(1):17–40
- Stajano F, Anderson R (1999) The resurrecting duckling: security issues for ad-hoc wireless networks. In: International workshop on security protocols

Secure Network Coding

- [Joint Network Coding and Opportunistic Forwarding](#)

Secure Routing in Cognitive Radio Vehicular Ad Hoc Networks

Ying He

Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

[A securing routing scheme for vehicular networks with cognitive radios](#); [Security in vehicular networks with cognitive radios](#)

Definitions

This topic mainly talks about a trust-based routing scheme for ad hoc vehicular networks with cognitive radios to effectively improve security for both spectrum sensing and data transmission processes.

Historical Background

In previous works, it is generally assumed that a spectrum-sensing attacker will only attack the spectrum sensing process in CR networks, and a data transmission attacker will only attack the data transmission process (Yu et al. 2009). However, it is highly possible for an intelligent attacker to attack both processes in CR-VANETs.

Vehicle Ad Hoc Networks

Recently, there is a phenomenal burst of interest in connected vehicles (CVs). CV systems use

connectivity (via advanced wireless communications) to enable vehicles, smart roadway infrastructure (SRI), and personal mobile devices to exchange information with each other and to provide road users with both safety and mobility advisories, warnings, and alerts (Yu 2016; He et al. 2018b). Vehicular ad hoc networks (VANETs) as the basic infrastructure can facilitate applications and services of CVs and intelligent transportation systems (Bravo-Torres et al. 2016; He et al. 2017). One of the main challenges in VANETs is the issue of limited resources, including radio spectrum resources, computation and storage resources, etc.

Cognitive Radio VANETs

For the limited radio spectrum issue, a promising solution is cognitive radio (CR) (Haykin 2005; Yu 2011), which has already extended to vehicular network environments (Haykin 2005). Generally, CR technology can help secondary users (SUs) that have no predefined spectrum for wireless communications to utilize the primary users (PUs) licensed spectrum. The condition is that the SUs should not interrupt the PUs normal communications. In other words, the interferences from the SUs are not allowed to produce a negative effect on PUs performance. In vehicular networks, CR technology can be used to mitigate the spectrum limitation issue, and when CR technology is introduced into VANETs, it is called CR-VANETs. In CR networks, there are two main processes: spectrum sensing and data transmission (Zhang et al. 2017). In the spectrum sensing process, mobile nodes detect the radio bands that are not utilized by PUs. In the data transmission process, data packets are transmitted using the detected idle radio bands.

Securing Routing in CR-VANETs

Both spectrum sensing and data transmission suffer many potential malicious attacks in CR-VANETs, such as incumbent emulation (IE) attack and spectrum sensing data falsification

(SSDF) attack (Sharifi and Niya 2016) in spectrum sensing, and packet dropping/modification attacks in data transmission. In an IE attack, an attacker can mimic a false signal of the PU in order to disturb the spectrum sensing process. In an SSDF attack, an attacker can disturb the SUs spectrum sensing results by disseminating fault spectrum sensing results deliberately in order to result in a wrong decision of the presence/absence of PUs. A packet dropping attack is also called as a black hole attack, which is a type of denial-of-service attacks (Djahel et al. 2011). In addition, packet modification of packets may have a significant impact on the performance of a network (Li and Trappe 2006).

Although some excellent works have been done to address spectrum sensing security and data transmission security in CR-VANETs, these two important areas have mostly been addressed separately in the existing works. In He et al. (2018a), an integrated framework is proposed to study trust management so as to enhance security for both spectrum sensing and data transmission processes in CR-VANETs. Additionally, a trust-based routing scheme for securing data transmission is presented.

Cross-References

- [Vehicle Ad-Hoc Networks: Services, Security, and Privacy](#)

References

- Bravo-Torres JF, Lpez-Nores M, Blanco-Fernndez Y, Pazos-Arias JJ, Ramos-Cabrera M, Gil-Solla A (2016) Optimizing reactive routing over virtual nodes in VANETs. *IEEE Trans Veh Technol* 65(4):2274–2294
- Djahel S, Nait-abdessalam F, Zhang Z (2011) Mitigating packet dropping problem in mobile ad hoc networks: proposals and challenges. *IEEE Commun Surv Tutor* 13(4):658–672
- Haykin S (2005) Cognitive radio: brain-empowered wireless communications. *IEEE J Sel Areas Commun* 23(2):201–220
- He Y, Yu FR, Zhao N, Leung VC, Yin H (2017) Software-defined networks with mobile edge computing and caching for smart cities: a big data deep reinforcement learning approach. *IEEE Commun Mag* 55(12):31–37
- He Y, Yu FR, Wei Z, Leung V (2018a) Trust management for secure cognitive radio vehicular ad hoc networks. *Ad Hoc Networks* To appear
- He Y, Zhao N, Yin H (2018b) Integrated networking, caching, and computing for connected vehicles: a deep reinforcement learning approach. *IEEE Trans Veh Technol* 67(1):44–55
- Li Q, Trappe W (2006) Reducing delay and enhancing dos resistance in multicast authentication through multigrade security. *IEEE Trans Inf Forensics Secur* 1(2):190–204
- Sharifi AA, Niya MJM (2016) Defense against SSDF attack in cognitive radio networks: attackaware collaborative spectrum sensing approach. *IEEE Commun Lett* 20(1):93–96
- Yu FR, Tang H, Huang M, Li Z, Mason PC (2009) Defense against spectrum sensing data falsification attacks in mobile ad hoc networks with cognitive radios. In: *IEEE Military Communications Conference*, pp 1–7
- Yu FR (2011) *Cognitive radio mobile ad hoc networks*. Springer, New York
- Yu FR (2016) Connected vehicles for intelligent transportation systems. *IEEE Trans Veh Technol* 65(6):3843–3844
- Zhang D, Chen Z, Ren J, Zhang N, Awad MK, Zhou H, Shen XS (2017) Energy-harvesting-aided spectrum sensing and data transmission in heterogeneous cognitive radio sensor network. *IEEE Trans Veh Technol* 66(1):831–843

Security

- [Interference Mitigation in Wireless Communications-Based Train Control \(CBTC\), Cognitive Control Approach](#)
- [Wireless Jamming Attack](#)

Security and Privacy in 4G/LTE Network

Nour Moustafa and Jiankun Hu
School of Engineering and Information
Technology, ADFA, Canberra, ACT, Australia

Synonyms

4G: Fourth generation; 4G/LTE technologies or 4G/LTE wireless networks; Advanced persistent threats(APT) or sophisticated attacks; LTE: Long-term evolution

Definitions

4G (Fourth Generation for wireless networks)

is a set of fourth-generation cellular data technologies, which enables multimedia communication and provides high data transfer rates.

LTE (Long-Term Evolution) denotes a standard for a smooth and efficient transition toward more advanced leading edge technologies for increasing the capacity and speed of wireless data networks.

TCP/IP (Transmission Control Protocol/Internet Protocol) refers to a set of protocols designed to construct a network of networks that a system uses for accessing the Internet/network.

APT (Advanced Persistent Threat) is a set of stealthy, advanced and continuous hacking processes, often launched and controlled by a hacker targeting a particular victim.

DoS (Denial of Service) is a cyberattack that attempts to disrupt services of a host or network.

Introduction

Over the last few decades, mobile systems have become essential for users to perform their daily tasks. This has led to the rapid evolution of wireless technologies, including the second generation (2G), third generation (3G), and fourth generation (4G) for mobile networks, to ubiquitously spread mobile telecommunication services (Seddigh et al. 2010). The 4G wireless technology has recently coined for improving broadband performance and allowing multimedia programs. Consequently, its architectures and standards have considerably enhanced to transfer higher data rates than 2G and 4G. Meanwhile, Long-Term Evolution (LTE) has evolved to become one of the effective technologies that accomplish the 4G wireless performance goals (Shaik et al. 2015).

Due to the high performances of 4G/LTE mobile devices, the LTE subscribers are expected to be about 3.16 billion by the end of 2018 (Statista 2018). There are various technological advances that 4G/LTE wireless networks provide when compared to earlier technologies. Firstly, 4G/LTE mobile systems work perfectly by utilizing the TCP/IP model. This, in fact, decreases financial and computational costs, where portable devices can connect to the Internet using an Internet Protocol (IP) without any constraints to previously closed cellular configurations. Nevertheless, with the wide variety of communication protocols included in the TCP/IP model, 4G/LTE wireless networks face multiple security and privacy issues (Seddigh et al. 2010; Shaik et al. 2015).

The key issues for securing 4G/LTE wireless networks can be summarized into three aspects. Firstly, mobile devices can flexibly access the Internet from any location and are therefore vulnerable to being hacked by different advanced persistent threats (APT). Secondly, while mobile IP-based systems are regularly updated with cryptographic and security mechanisms, there is an effect on their performance and traffic processing capacity that requires secure and upgraded wireless standards and architectures. Finally, although vendors are producing new generations of 4G/LTE technologies, they do not regularly develop new standards to mitigate vulnerabilities and deter growing cyber APT (Seddigh et al. 2010; Shaik et al. 2015; Li et al. 2018).

Background

This section discusses the background of 4G wireless standards and LTE architectures. Security controls of 4G/LTE architecture are also explained.

4G Wireless Standards

The International Telecommunications Union (ITU) declared an International Mobile Telecommunications-Advanced (IMT-Advanced) standard for 4G wireless networks. This standard

Security and Privacy in 4G/LTE Network, Table 1 Characteristics of 2G, 3G, and 4G technologies

Features	2G	3G	4G
Standards	GSM, iDEN,D-MPS	WCDMA, CDMA 2000	Single unified standard, ITU IMT-Advanced
Data rates	14.4 kbps	2 Mbps	100 Mbps
Services	Digital voice, Short Messaging	high-quality audio, video, data	Dynamic information access with higher multimedia quality, wearable devices
Technology	Digital cellular	Broad bandwidth CDMA, IP technology	Unified IP, seamless combination of broadband, LAN/WAN/PAN, WLAN
Core network	PSTN	Packet Network	Internet
Multiplexing	TDMA, CDMA	CDMA	CDMA

provides the specifications of radio access and core 4G wireless networks. The 4G wireless technology includes the following criteria: (1) high data rate, which is 100 Mbps for mobile devices and 1 Gbps for computer devices, (2) high quality of service (QoS) and capacity, and (3) high network speeds and coverage (Seddigh et al. 2010; Rao et al. 2017). The characteristics of 4G technologies compared with 2G and 3G are listed in Table 1 (Fagbohun 2014).

The bandwidth efficiency and allocation schemes are two important requirements that should be considered while designing 4G standards (Seddigh et al. 2010). In 4G wireless, voice/video multimedia is transited using the network protocols of the TCP/IP model. Therefore, the ITU-IMT-Advanced standard for 4G wireless technology should be configured to be compatible with the protocols and services of the TCP/IP model. Several 4G wireless standards, in particular, LTE and Mobile WiMAX, have been developed to meet the IMT-Advanced requirements and provide broadband wireless connections for mobile devices (Seddigh et al. 2010; Rao et al. 2017).

LTE Architecture

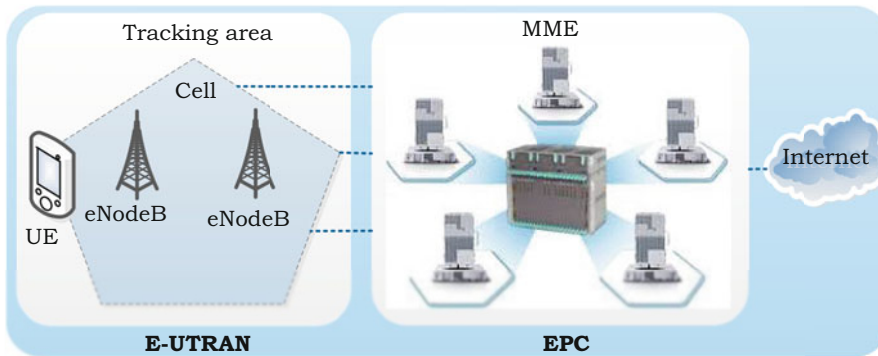
A LTE architecture includes the modules needed to install network protocols between base stations and mobile systems. As presented in Fig. 1, the architecture involves three modules: User Equipment (UE), Evolved Universal Terrestrial Radio Access Network (E-UTRAN), and Evolved Packet Core (EPC) (Seddigh et al. 2010; Shaik et al. 2015). The UE, for example, laptops

or smartphones, can link to the wireless network across the evolved NodeB (eNodeB) using the E-UTRAN base stations. The eNodeB utilizes some access network protocols for exchanging messages with the UE. The E-UTRAN links to the EPC which is an IP-based infrastructure, while the EPC links to the provider of the wireline IP network.

The 4G/LTE network architecture has some enhancements compared to 3G wireless (Shaik et al. 2015). Firstly, it has two types of network elements (NEs): (1) the eNodeB that is an improved base station and (2) the Access Gateway (AGW) that integrates all the functions, specifically Mobility Management Entity (MME), needed for the EPC. The MME can control the UE identification, as well as processing security authentication and mobility. LTE can support a meshed structure that improves wireless network performance, for example, an eNodeB can connect with several AGWs. Finally, as the architecture is compatible with the TCP/IP model, traffic packets at any UE can be handled using the AGW and eNodeB with different IP-based devices, such as routers.

4G/LTE Security Controls

Abstraction layers are inserted in the 4G/LTE architecture in terms of the unique identifiers (IDs) for smartphones (i.e., UEs). A temporary unique ID is used on the SIM card to prevent attackers from stealing identifiers. Another technique for improving 4G security is adding protected singling between the UE and MME (Seddigh et al. 2010; Mohapatra et al. 2015). Security



Security and Privacy in 4G/LTE Network, Fig. 1 LTE network architecture

mechanisms are utilized to secure the connections between 4G networks and secure non-4G networks using key management authentication protocols. Although several security controls are used for 4G/LTE wireless technology, its design, which is based on an open-IP architecture, and the sophistication of APT hackers make security and privacy of 4G/LTE systems challenging.

4G/LTE Security Requirements

In order to secure mobile devices that use 4G/LTE wireless technologies, there should be protection for the connections between the UEs and MMEs and between elements in the wireline networks and mobile stations. For satisfying these requirements, the 4G/LTE security is significantly improved by adding (1) advanced key hierarchy, (2) protracted authentication and key agreement, and (3) additional interworking security for the NEs (Mohapatra et al. 2015). The requirements are classified into key building blocks and LTE end-to-end security (Seddigh et al. 2010), as explained below.

- **Key building blocks** include the following elements:

- **Key security and hierarchy**

LTE has five key strategies used for connections of the EPS and E-UTRAN. The keys are declared as follows: (1) KANS encryption and integrity keys are used to protect non-access stratum (NAS)

traffic between the UE and MME, (2) a KUP encryption is used to encrypt traffic between the UE and eNodeB, and (3) KPRC encryption and integrity keys are used to secure the Radio Resource Control (RRC) between the UE and eNodeB.

- **Key management**

LTE key management comprises three functions: key establishment, distribution, and generation. It is essential that 4G/LTE wireless technology has key management mechanisms that prevent stealing keys, as mobile devices with IP-based infrastructure can frequently access different wireless networks. An Authentication and Key Agreement (AKA) process is utilized for establishing and verifying keys in 4G/LTE systems.

- **Authentication, encryption, and integrity protection**

LTE depends on using regular updating of the authentication process by exchanging sequence numbers in the messages of encryption mechanisms. The IPsec protocol and tunnels are also used for asserting the confidentiality of users' data while transmitting traffic between LTE nodes.

- **Unique user identifiers**

LTE has several user identifier mechanisms that thwart attackers from learning mobile user identities; therefore, attackers cannot track user profiles or launch denial of service (DoS) attacks against users. The identifier mechanisms

contain the following: (1) international mobile equipment identifier (IMEI) which is a permanent unique identifier for each mobile, (2) M-TMSI which is a temporary identifier that defines the UE inside the MME, and (3) cell radio network temporary identifier (C-RNTI) which is a unique and temporary UE identity when a UE is connected with a cell.

- **LTE end-to-end security** involves the following elements:

- **Authentication and Key Agreement (AKA)**

The foundation of LTE security is authenticating the UEs and wireless networks. This can be accomplished using the AKA process which asserts that the serving network authenticates the identity of a user and the UE certifies the network signature. The AKA creates encryption and integrity keys applied for originating various session keys for ensuring the 4G/LTE security and privacy.

- **Confidentiality and integrity of signaling**

Security of network access control planes is achieved when the RCC and NAS layer signaling is encrypted and integrity protected. Ciphering and integrity protection of LTE RRC signaling is executed at the packet data convergence protocol (PDCP) layer, whereas the NAS layer attains the protection by encrypting the NAS-level signaling. This protection cannot be uniquely performed for each UE connection, but it runs across trusted connections between AGW and eNodeB.

- **User plane confidentiality**

LTE has a security feature for user plane via encrypting data/voice between the UE and eNodeB. Encryption is executed at the IP layer by utilizing IPsec-based tunnels between AGW and eNodeB, but no integrity protection is offered for the user plane due to performance and efficiency considerations. The PDCP layer is used for enabling encrypting/decrypting the user

plane while transmitting traffic between the eNodeB and UE.

Cyberattacks and Countermeasure Techniques

4G/LTE wireless technology faces different types of cyberattacks that could affect integrity, privacy, availability, and authentication, as described below.

- **Privacy attacks**

Attacks against the privacy of mobile users' data attempt to expose sensitive data/multimedia of users. A man-in-the-middle (MITM) attack is the most serious privacy attacks in wireless networks that depend on a false base station attack when anomalous third-party masquerades its base transceiver station (Mohapatra et al. 2015). Privacy-preserving authentication and encryption mechanisms have been widely used to protect wireless networks against the MITM attacks (Ferrag et al. 2017; Deebak et al. 2016).

- **Integrity attacks**

Attacks against integrity attempt to modify exchanging data between the 4G access points and mobile users. Cloning attacks based on the MITM and message modification scenarios are the major integrity attacks that alter mobile user information. Authentication and privacy-preserving mechanisms with hash functions have been broadly used for securing 4G wireless networks against integrity attacks (Ferrag et al. 2017; Hasan et al. 2017).

- **Authentication attacks**

Attacks against authentication attempt to disturb the client-to-server and/or server-to-client authentication process. The password reuse, brute force, password stealing, and dictionary attacks are popular wireless hacking schemes that interrupt the password-based authentication. In the hacking schemes, an attacker can pretend to be a legal user

and try to log in to a server by guessing various words as a password from a dictionary. Encryption and authentication techniques have been utilized for preventing such kind of attacks from 4G wireless networks (Seddigh et al. 2010; Ferrag et al. 2017).

• **Availability attacks**

Attacks against availability try to make services unavailable, such as the service of data routing (Ferrag et al. 2017). The first in first out (FIFO) and DoS attacks can be launched by flooding massive malicious actions to 4G wireless victims for disrupting their computational resources. Firewall and intrusion detection systems have been usually used for defending against these attacks.

4G/LTE Challenges and Future Trends

Despite a plethora of research and technical studies that have been conducted for securing 4G/LTE wireless networks, there are several challenges that should be the focus of researchers in future that are discussed below, besides a summary in Table 2.

- Designing a flexible and scalable 4G/LTE architecture that can address security and privacy issues is an arduous task. There are multiple devices and systems that are usually

connected with 4G networks that result in vulnerabilities and loopholes in networks.

- Discovering DoS attacks that attempt to violate 4G wireless networks, as hackers frequently establish new sophisticated variants against eNodeB, UE, and discontinuous reception services.
- Location tracking denotes tracing the UE presence in a specific cell(s). While many portable devices could link to a 4G/LTE wireless network, ensuring that location tracks of the devices are not breached is still a challenging issue, due to the considerations of operability and scalability.
- The utilization of an effective 4G wireless Software Dened Network (SDN) is a challenge. More specifically, there are technical gaps in the network scalability, security, and privacy issues with the SDN.
- Trusted connections through 4G networks in the existence of eavesdroppers are the issues. Especially, when 4G wireless technology is used in the Internet of Things, it requires new cryptographic mechanisms that provide protection and integrity for smartphones and computer systems.
- Instead of individual security techniques, a systematic security and privacy protection strategies are required for 4G/LTE wireless connections while connecting with cloud and edge computing paradigms. This will provide

Security and Privacy in 4G/LTE Network, Table 2 Challenges and security and privacy methods of 4G/LTE technology

Challenges	Cyber-attacks	Security and privacy methods
A resilient 4G/LTE architecture	Privacy attacks: replay, MITM, impersonation, collaborated, tracing, spooing, privacy violation, masquerade	Privacy-preservation, authentication and encryption mechanisms
Tracking locations of devices	Integrity attacks: cloning, spam, message blocking, message modification attack, message, insertion, tampering	Hashing and encryption, and authentication and privacy-preserving methods
An effective 4G/LTE wireless Software Defined Network (SDN)	Availability attacks: FIFO, redirection, physical attack, skimming, and free-riding	Firewall systems, signature-based and anomaly-based systems
Collaborative 4G/LTE security and privacy approaches operate on cloud and edge paradigms	Authentication attacks: password reuse, password stealing, dictionary, brute force, desynchronization, forgery attack, collision, stolen smart card	Encryption and authentication techniques

valid security mechanisms, for example, trust models, device security, and data assurance techniques.

Key Applications

The 4G/LTE wireless technology has emerged for enhancing broadband performance and permitting multimedia applications. The technology has been used for the Internet of Things (IoT) for connecting Machine-to-Machine (M2M) systems and devices such as the In-Vehicle Multi-Carrier Router.

Cross-References

- [4G/LTE Technologies or 4G/LTE Wireless Networks](#)
- [Wireless Data Security](#)

References

- Deebak BD, Muthaiah R, Thenmozhi K, Swaminathan PI (2016) Analyzing the mutual authenticated session key in ip multimedia server-client systems for 4G networks. *Turk J Electr Eng Comput Sci* 24(4):3158–3177
- Fagbohun O (2014) Comparative studies on 3G, 4G and 5G wireless technology. *IOSR J Electr Commun Eng* 9(3):88–94
- Ferrag MA, Maglaras L, Argyriou A, Kosmanos D, Janicke H (2017) Security for 4G and 5G cellular networks: a survey of existing authentication and privacy-preserving schemes. *J Netw Comput Appl* 101:55–82
- Hasan K, Shetty S, Oyedare T (2017) Cross layer attacks on GSM mobile networks using software defined radios. In: 2017 14th IEEE annual consumer communications & networking conference (CCNC). IEEE, Las Vegas, NV, pp 357–360
- Li S, Da Xu L, Zhao S (2018) 5G internet of things: a survey. *J Ind Inform Integr* 10:1–9
- Mohapatra SK, Swain BR, Das P (2015) Comprehensive survey of possible security issues on 4G networks. *Int J Netw Secur Appl* 7(2):61
- Rao KV, Rao OK, Sainath MN (2017) A study on broad spectrum of various wireless technologies. *Int J Innov Adv Comput Sci* 6(12):354–360
- Seddigh N, Nandy B, Makkar R, Beaumont JF (2010) Security advances and challenges in 4G wireless networks. In: 2010 eighth annual international conference on privacy security and trust (PST). IEEE, Ottawa, ON, pp 62–71
- Shaik A, Borgaonkar R, Asokan N, Niemi V, Seifert JP (2015) Practical attacks against privacy and availability in 4G/LTE mobile communication systems. arXiv preprint arXiv:151007563
- Statista (2018) LTE subscribers. <https://www.statista.com/statistics/206615/>

Security in Delay-Tolerant Networks

Qiuming Liu and He Xiao
Department of Software Engineering, Jiangxi
University of Science and Technology,
Nanchang, Jiangxi, China

Synonyms

[Authentication in delay-/disruption-tolerant networks](#); [Security in disruption-tolerant networks](#)

Definition

Authentication and access control policies to delay-/disruption-tolerant networks with limited bandwidth, relatively high bit error rates, and periods lacking connectivity interworking environment.

Historical Background

The history of delay-tolerant networking began as projects under United States government grants relating to the necessity of networking technologies that can sustain the significant delays and packet corruption of space travel. Initially, these projects looked only short-range communication between manned missions to the moon and back, but the field quickly expanded into an entire sub-field of DTNs that created the technological advances to allow for the Interplanetary Internet.

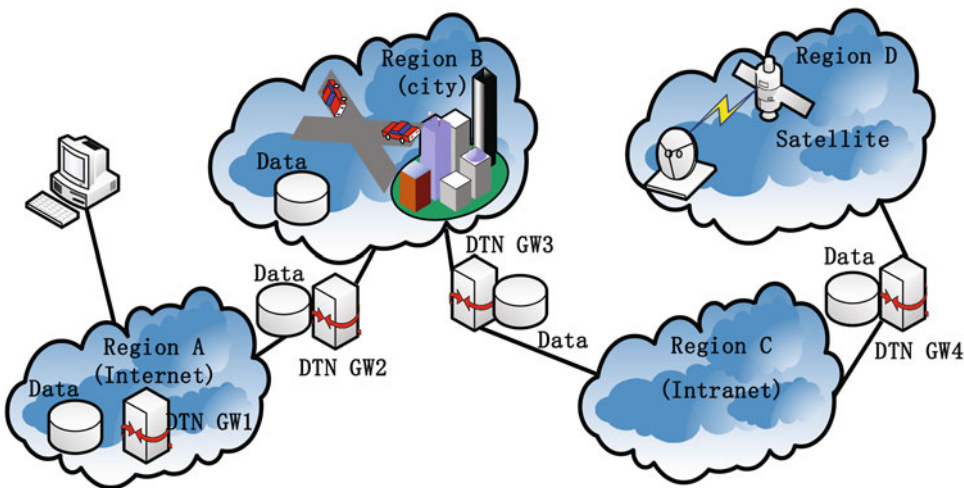
Introduction

The traditional transfer control protocol/Internet protocol (TCP/IP) plays an important role in

current Internet service (Akyildiz et al. 2004). Although not clearly stated, the existing Internet service is based on a number of key assumptions, such as an end-to-end link always exists between the source and its destination, the maximum round-trip time is not excessive, and the packet loss probability is small. However, a class of challenged networks would violate one or more of the assumptions, which will lead to this kind of network that may not be well served by the current end-to-end TCP/IP protocol. The architecture involved here is called delay-/disruption-tolerant networks (DTNs) (Fall 2003). The DTNs emerge in many circumstances, such as mobile ad hoc networks (MANET) (Zhang 2007), vehicle networks (VANET) (Pereira et al. 2012), satellite networks (Wang et al. 2013), etc. Different from the traditional network model, DTNs may have frequency disruption and long delay and no end-to-end link simultaneously, and the network model is constraint by limited storage/processing. As illustrated in Fig. 1, the DTN architecture includes the concepts of regions and DTN gateways (GWs). In Fig. 1, there are four network regions (A, B, C, D). Region B includes a DTN GW which locates on vehicles that cycle between DTN GW 2 and 3. Region D includes a low-

earth-orbiting (LEO) satellite link that also provides periodic connectivity (albeit perhaps more regular than the vehicle which may be subject to vehicular congestion or other delays).

The speciality of DTNs would result in TCP that cannot satisfy the applications over DTNs. To deal with the new challenges in DTNs, a new bundle protocol is overlaid on the transfer layer. As an “overlay” architecture, DTNs are intended to operate above the existing protocol stacks in various network architectures and provide a store-and-forward GW function between them when a node physically touches two or more dissimilar networks. For example, within the Internet, the overlay may operate over TCP/IP, and for delay-tolerant sensor/actuator networks, it may provide interconnection with some sensor transport protocol. Each of these networking environments has its own specialized protocol stacks and naming semantics developed for their particular application domain. Achieving interoperability between them is accomplished by special DTN GW located at their interconnection points. However, due to the speciality of DTNs, the security mechanism is different from that of traditional Internet networks. The main venue for DTN related security work is the Internet Research Task Forces (IRTF) (IRTF



Security in Delay-Tolerant Networks, Fig. 1 A typical DTN network architecture

2014) and Delay Tolerant Networking Research Group (DTNRG) (DTNRG 2009). The primary objective here is to overview those defining security services for DTNs.

Security in DTNs

In DTNs, the transmission media is usually oversubscribed and link capacity is a precious resource (Farrell et al. 2009). Thus, it is necessary to design some authentication and access control mechanism to protect the service of data forwarding, especially for the critical points in the network. Since there are multiple classes of services, the security mechanism considering endpoints is not very attractive. The reason is that end-to-end link mechanism requires some exchange of keys, which would be undesirable for high-delay and disconnection-prone networks. Moreover, it is undesirable to carry unwanted traffic all the way to its destination before an authentication and access control check is performed (Zhu et al. 2009).

Bundle Protocol Security

The unique data security in DTNs is focused on bundle layer, including faking bundle source, distorting the destination and control domain, modifying payload, etc. Such would consume the network resource and influence the confidentiality and reliability of payload in bundle layer. The regular way to solve the security of bundle protocol is encryption. However, due to the feature of DTNs, traditional encryption cannot apply, since it requires multiple interoperable and much resource consumption (Menesidou and Katos 2017).

To utilize the link resource and eliminate the large delay in DTNs, bundle protocol adopts opportunity routing to deliver data block. Different from that of the Internet, DTNs cannot authenticate the block in time, since there exists large delay, such that it is easily attacked by malicious nodes. Therefore, the bundle protocol needs to support a better block authentication.

Addressing security issues has been a major focus of the bundle protocol. Security concerns

for delay-tolerant networks vary depending on the environment and application, though authentication and privacy are often critical. These security guarantees are difficult to establish in a network without persistent connectivity because the network hinders complicated cryptographic protocols and key exchange, and each device must identify other intermittently visible devices. Solutions have typically been modified from mobile ad hoc network and distributed security research, such as the use of physical layer security.

Key Applications

The security of DTNs are involved in mobile ad hoc networks, vehicle networks (VANET), satellite networks, airborne networks, etc. Figure 1 illustrates a typical DTN, and there is an Internet connection available in a nearby town and a transport medium from the rural area to the town in the form of a vehicle, such as a bus or a car. There are two types of end users, mobile users, who use their own personal devices to connect directly to gateways. The bundle protocol security communication architecture targets mainly mobile users. Achieving security in such disconnected networks is a demanding task, but it is necessary in hostile environments with malicious attackers or even just passive listeners. In a hostile environment, secure DTN communication can provide an efficient way to let informers transfer information while hiding their identity from an enemy. Therefore, the utility of bundle protocol security in the DTNs is greatly expanded when the DTNs provide end-to-end security.

References

- Akyildiz IF, Akan OB, Chen C et al (2004) The state of the art in interplanetary internet[J]. *IEEE Commun Mag* 42(7):108–118
- Delay Tolerant Networking Research Group (2009) <http://www.dtnrg.org/>
- Fall K (2003) A delay-tolerant network for challenged internets. *ACM Sigcomm, Karlsruhe*, pp 27–34

- Farrell S, Symington SF, Weiss H, Lovell P (2009) Delay-tolerant networking security overview. DTN Research Group (Internet-Draft). Retrieved from <https://tools.ietf.org/html/draft-irtf-dtnrg-sec-overview-06>
- Internet Research Task Force (2014) <http://www.irtf.org/>
- Menesidou SA, Katos V (2017) Authenticated Key Exchange (AKE) in delay tolerant networks. IFIP international information security conference. Springer, Berlin/Heidelberg, pp 49–60
- Pereira PR, Casaca A, Rodrigues JJPC et al (2012) From delay-tolerant networks to vehicular delay-tolerant networks. *IEEE Commun Surveys Tuts* 14(4): 1166–1182
- Wang R, Wei Z, Zhang Q et al (2013) LTP aggregation of DTN bundles in space communications. *IEEE Trans Aerosp Electron Syst* 49(3):1677–1691
- Zhang Z (2007) Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges[J]. *IEEE Commun Surveys Tuts* 8(1):24–37
- Zhu H, Lin X, Lu R et al (2009) SMART: a secure multi-layer credit-based incentive scheme for delay-tolerant networks. *IEEE Trans Veh Technol* 58(8):4628–4639

Security in Disruption-Tolerant Networks

- [Security in Delay-Tolerant Networks](#)

Security in Vehicular Networks with Cognitive Radios

- [Secure Routing in Cognitive Radio Vehicular Ad Hoc Networks](#)

Sensing Augmented

- [Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement](#)

Sharing Economics

Costas Courcoubetis¹ and Richard Weber²

¹Engineering Systems and Design, Singapore University of Technology and Design, Singapore, Singapore

²Statistical Laboratory, University of Cambridge, Cambridge, UK

Synonyms

[Collaborative consumption](#); [Crowdsourced wireless networks](#); [Infrastructure sharing](#), [Wi-Fi sharing](#); [Network sharing](#); [Wireless community networks](#)

Definition

A sharing economy is an economic model in which individuals (businesses or consumers) are able to borrow or rent assets owned by others. This model is usually successful when the price of a particular asset is high, and it is not fully utilized by its owner. Collaborative consumption is an example of a sharing economy in which there is shared use of assets or services by a group. It differs from standard commercial consumption in that the cost of purchasing the assets or services is not borne by one individual but instead is divided across a larger group, with the asset costs usually being recovered through renting or exchanging. In communication networks economic models of sharing are being used to share expensive infrastructure among network operators or to create a whole access network via individuals' Wi-Fi sharing.

A key idea behind such sharing models is the paradigm shift away from the exclusive ownership and use of assets or resources to one of shared use. Value is derived from the fact that most assets are acquired to satisfy peak demand but are otherwise poorly utilized. So sharing has the benefit of increasing access to resources while reducing investments in assets. This movement takes advantage of innovative new ways

of peer-to-peer (P2P) sharing that are voluntary and enabled by Internet-based exchange markets and mediation platforms. Several successful businesses like Uber and Airbnb in the USA and elsewhere provide a proof-of-concept and evidence for the viability of the sharing economy.

A key ingredient of a P2P sharing economy application is the Internet-based platform that implements the underlying economic model. Its success depends on reducing the various transaction costs, creating not just a marketplace but also a market for the given service, and becoming a trusted third party that enforces the rules of the service and checks for quality.

In the case of telecom services, one can broadly distinguish two types of sharing models. The first type is a business-to-business (B2B) case, where basic network infrastructure is shared among competing providers in order to minimize their costs. The economic drive is to achieve large economies of scale. The second type concerns P2P sharing where the network infrastructure and the corresponding service is provided by individuals instead of companies, as in the case of Wi-Fi sharing to produce ubiquitous Internet access. In this case the success of the model depends on the large positive externalities of offering network access as a public good.

Historical Background

The Case of B2B

The telecommunication industry enjoys large economies of scale, and it is typical that telecom operators must make considerable investment in technology and infrastructure and upgrade this every few years. Consequently, infrastructure sharing, as it limits duplication, is becoming a successful business paradigm in the industry. In many cases competitors are becoming partners in order to lower their continuing investments and the new technology upgrades. The degree and method of infrastructure sharing vary in different countries depending on regulatory and competitive conditions. From the market efficiency point of view, infrastructure sharing can significantly facilitate entry for new players

by decreasing the level of required investments, with great impact on competition. The market becomes more attractive to new players, with a lower barrier to entry and reduced business risk. Infrastructure sharing allows operators to focus more upon improved innovation and better commercial offerings, increasing consumer surplus (see OECD 2015).

There are two basic categories of sharing mobile infrastructure: passive and active. Passive sharing refers to the sharing of physical space (i.e., in buildings, sites, and masts) where networks remain separate and run their own equipment. In active sharing, active elements of a mobile network are shared, such as antennas, entire base stations, or even functions of the core network. A popular form of active sharing is mobile roaming. An operator uses the infrastructure of another network in geographic areas where it has no infrastructure of its own.

However, the decisions faced by mobile network operators on which resources they might share, and how they should compete with operators with whom they do share facilities, are subtle and might depend on the type of contracts that will be in place. There is always the risk of a competitor that shares infrastructure to learn strategic information like traffic density, network load, etc. Also sharing might impact an operator's capability for competitive differentiation, its freedom in making decisions for new technology investments, and its operational practices. Regulators may be concerned that network sharing may have negative impact upon competition. Two or more operators could share resources in such a way creating anti-competitive discrimination against other operators.

The Case of P2P

Wi-Fi signals of access points (APs) set up by individuals for their own use already pervade many cities, and all together cover a significant fraction of the city's geographical area. In most cases such a network cell installed by its owner covers a greater area than originally intended and hence is accessible to other individuals that are nearby. Given how easy it is to gain access to an AP once a potential user is within its

coverage area, and leaving aside the obvious security issues involved, it is sensible that individuals could share such infrastructure among themselves to achieve ubiquitous Internet access. Sharing comes as a natural idea since Wi-Fi technology provides large amounts of bandwidth that is mostly underutilized by its original user both at the access level and also at the backhaul. So the important questions are (i) How does a “primary” user gain by sharing her AP with other “secondary” users and (ii) make sure that she does not lose by doing so (decrease of network speed, security risks).

The key idea is to offer the right incentives to individuals to share their owned private Wi-Fi APs with each other, i.e., to crowdsource the public network coverage of large geographic areas using private APs. Users can benefit from joining such a community network by using other users APs when traveling to the corresponding locations. Clearly, the success of such a crowdsourcing model largely depends on the degree of participation and contribution of many individual Wi-Fi owners and requires a careful design of the economic incentives. Key issues for success are (i) the ability of the local AP to isolate secondary user traffic from primary user traffic for performance and security reasons, (ii) the distributed management of such a large infrastructure, (iii) the size of the anticipated network since a larger network is more useful to its users, and (iv) pricing and economic incentives in general. It should be stressed that although technology can solve the first two, pricing is rather open and can lead to different business models. For example, the platform could operate in a tit-for-tat fashion, blocking users whose AP is turned off to secondary users. Or subsidize users for the cost of the new AP equipment or for part of their broadband connection to the ISP. Profits could be made by charging users from other communities that like to use the platform or from mobile network operators (MNOs) by allowing them to offload data of their customers.

A prominent example of a commercial wireless community networks is Fon (see [FON web-site](#)) which has more than 20 million member Wi-Fi APs around the world. The original idea was to

create an ubiquitous network for Wi-Fi access for its customers. To become a Fon member, users had to buy a special router called a “Fonera.” With this router, they agreed to share part of their bandwidth as a second Wi-Fi signal, in exchange for the right to use other members’ Fon APs. More precisely, the operator incentivizes Wi-Fi sharing using two incentive schemes corresponding to two types of membership: as a “Linus,” a user can get free Wi-Fi access within the community network coverage by sharing her Wi-Fi for free, and as a “Bill” a user can earn an additional fraction of the revenue generated from roaming traffic after subtracting certain costs. As an “Alien” a user who does not own a Wi-Fi AP can still access the Fon network by purchasing Wi-Fi passes from the operator.

Foundation

Underlying the B2B and the P2P sharing models are some basic economic models.

In the B2B model, partnering businesses jointly pay for a common infrastructure, which is useful to them all, but not necessarily to the same extent. If all partners are given equal access and are expected to share equally the cost, then the amounts that partners are willing to contribute to such cost can be less than optimum and lead to a “poor infrastructure” system and also free-riding by partners who ought to be paying more. Free-riding takes can place when a participant can use infrastructure built by others while contributing very little to its total cost. When all participants expect to be able to free-ride then a socially inefficient choice of poor infrastructure system can occur.

A way to remedy that is to request participants to contribute some minimum amount or incentivize them to contribute more by differentiating the quality of service obtained by using the infrastructure, i.e., by relating the amount of cost that a partner bears to priorities in accessing the infrastructure, etc. These issues are illustrated in the following simple model of economic agents that build a common infrastructure and derive benefit from its usage.

Suppose that a shared infrastructure is to be built, of quality or quantity Q . This can be modelled as a public good since its larger size benefits all agents that have access to it. The social welfare obtained by a population of n agents is then

$$\sum_{i=1}^n \theta_i u(Q) - c(Q).$$

Here u is an increasing concave function assumed known, and θ_i is a parameter which expresses the importance that agent i attaches to using the system. The optimal choice of Q will depend on all the values of the $\theta_1, \dots, \theta_n$, which are the private information of the agents: i.e., θ_i is known only to agent i , though its a priori distribution is known. One must ask agents to reveal their θ_i and then choose Q and impose agent-specific payments, each of which depends on all of $\theta_1, \dots, \theta_n$. These payments should cover the cost $c(Q)$. However, the agents, knowing how the above will be done, may act strategically, i.e., agent i may choose to declare a non-truthful value $\hat{\theta}_i \ll \theta_i$ in order to be charged a smaller fraction of the total cost. In doing so she assumes that all other agents j will declare their true values of θ_j , and hence the system will be built at a large scale, not being much influenced by her non-truthful declaration. Note that this strategy makes more sense as the number of participants gets larger. If all agents declare small $\hat{\theta}_i$ s, the corresponding optimal choice of Q will also be small.

The so-called mechanism design problem is to devise rules that will obtain the best outcome subject to agents acting strategically. When n is small, this is exceedingly difficult. However, as described in Courcoubetis and Weber (2012), there is, when n is large, a nearly optimal design in which, by knowing only n and the expected average of the θ_i 's, one can set $Q = Q^*$ and impose on each agent the same fixed fee p for participation. Agent's will self-select: i.e., agent i chooses to participate only if $\theta_i u(Q^*) - p > 0$. By extending this simple model, one can further investigate the effect of offering differentiated quality of service in proportion to the cost covered by each agent.

In P2P, agents contribute in kind. Each agent benefits from the contribution of others resources but pays only the cost of its own contribution. Consider, for example, a population of agents; suppose agent j_i makes available some shared Wi-Fi service which other agents can connect to and use when roaming in her locale j . She provides this "connectivity service" in quality Q_{j_i} (such as availability or download speed). The total available for the community in locale j is $Q_j = Q_{j_1} + \dots + Q_{j_{n_j}}$, summed over the n_j agents who provide shared Wi-Fi in locale j (though as concerns agent j_i , this should omit her own contribution Q_{j_i} to the community Q_j , for while she benefits from the contributions of other agents in locale j she has anyway Wi-Fi connectivity in her own premises.) Agent j_i incurs a cost in providing Q_{j_i} which is a function of both Q_{j_i} and the load the other agents place on her networking equipment, say $c(Q_{j_i}, \cdot)$. By using the Wi-Fi of others, agent j_i obtains for herself net benefit of the form

$$\theta_{j_i} \sum_{k=1}^L u(Q_k) - c(Q_{j_i}, \cdot), \quad (1)$$

where $\theta_{j_i} u(Q_k)$ measures the utility she receives due to roaming in locale k . Suppose the form of u is known and the same for all agents and locales, but θ_{j_i} is information private to agent j_i . Now there can be a free-rider problem, in that the user j_i might do best for herself by setting $Q_{j_i} = 0$. Of course there is an easy remedy: simply require her to provide Wi-Fi of at least some specified amount $Q_{j_i}^*$, or otherwise she is barred from sharing the wireless of others when she roams.

When the aim is to maximize total social welfare (which is the sum of (1) taken over all agents), then optimal values of the $Q_{j_i}^*$ will depend on the values of all the θ_{j_i} , which we have assumed are not known to the system designer. However, when the total number of agents is large, a scheme can be designed in which each agent in locale j is asked to contribute some minimum quality $\bar{Q}_j$. This will obtain nearly the same optimum social welfare as when all θ_i are known and appropriate optimal $Q_{j_i}^*$ are required of the agents. Some agents, who know their θ_{j_i}

to be small, will choose not to participate in such a system. But those who do participate will not be free-riding, and the system will dimension to the optimal scale (see Courcoubetis and Weber 2004). Similar models have been analyzed in Manshaei et al. (2008), Afrasiabi and Guerin (2012) and Ma et al. (2017).

Cross-References

- Auction
- Behavioral Economics
- Budget Feasible Mechanisms
- Collusion
- Dynamic Pricing
- Efficiency and Pareto Optimality
- Fairness
- Market Entry
- Matching for Cooperative Spectrum Sharing
- Oligopoly Pricing in Wireless Networks
- Preferences and Utility Functions
- Sharing Economics

References

- Afrasiabi MH, Guerin R (2012) Pricing strategies for user-provided connectivity services. In: Proceedings of IEEE INFOCOM mini-conference
- Courcoubetis C, Weber RR (2004) Asymptotics for provisioning problems of peering wireless LANs with a large number of participants. In: Proceedings of WiOpt04: modelling and optimization in mobile, ad hoc and wireless networks
- Courcoubetis C, Weber RR (2012) Economic issues in shared infrastructures. *IEEE/ACM Trans Netw* 20:594–608
- FON website. <https://corp.fon.com>
- Ma Q, Gao L, Liu JF, Huang J (2017) Economic analysis of crowdsourced wireless community networks. *IEEE Trans Mob Comput* 16:1856–1869
- Manshaei MH et al (2008) On wireless social community networks. In: Proceedings of IEEE INFOCOM 2008, pp 2225–2233
- OECD (2015) Wireless market structures and network sharing. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP/CISP\(2014\)2/FINAL&docLanguage=En](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP/CISP(2014)2/FINAL&docLanguage=En)

Shark Turning Classification

- [Turning Detection in Sandbar Sharks Through Accelerometer Data](#)

Shark Turning Prediction

- [Turning Detection in Sandbar Sharks Through Accelerometer Data](#)

Sidelink Transmissions and Proximity Services in LTE-A and Beyond

Shao-Yu Lien¹ and Der-Jiunn Deng²

¹National Chung Cheng University, Chia-Yi, Taiwan

²Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua City, Taiwan

Synonyms

[Device-to-device \(D2D\) communications](#); [Direct transmissions](#)

Definition

Sidelink in LTE-A denotes a transmission direction from a user equipment (UE) to another UE(s), in contrast to downlink from an eNB to a UE and uplink from a UE to an eNB. In sidelink transmissions, if a UE has data to be transmitted to another UE(s), a UE can directly transmit data to the destination UE without signal relay via an eNB. Consequently, sidelink transmissions can only take place among two or more UEs within the signal coverage of each other, and such a framework is also known as proximity services (ProSe) or direct transmissions.

Historical Background

Since 2012, the First Responder Network Authority (FirstNet) of the United States (US) Department of Commerce (DoC) is tasked with creating a nationwide network to provide wireless services for public safety agencies across the United States. However, as the costs to develop and to maintain a mobile network particularly for public safety are drastically increasing, this obstacle consequently drives the exploitation of existing telecommunication infrastructures. Considering the economic scale of LTE/LTE-A and the capabilities to provide multi-megabit per second data rates, 3GPP Release 12 is thus endorsed as the next-generation nationwide public safety network in the United States. Using dedicated 700 MHz spectrum allocated by the Federal Communications Commission (FCC), emergency responders such as police, fire brigade, and ambulance are able to receive immediate wireless services when public safety events (such as rescue, conflagration, riot, accident, etc.) occur. However, although most requirements for public safety can be supported by LTE/LTE-A Ferrus et al. (2013), there is a critical concern. When a part of or all eNBs are unavailable due to natural disasters or malicious attacks, wireless services still need to be provided at least locally for a group of users. For this particular requirement needing to be supported by a public safety network, two key technologies are supported by 3GPP Release 12: D2D proximity services (ProSe) and group communications (3GPP TR 36.843 V12.0.0 2014; Lien et al. 2016a).

Enabling UEs to directly exchange data without signal relays via eNBs, D2D communications have been well studied in the literature to provide considerable benefits, including significant improvements in spectrum (reuse) efficiency (Yu et al. 2011; Peng et al. 2014), transmission data rate using higher-order modulation and coding schemes (MCSs), power saving, and latency performance (Liu et al. 2015; Peng et al. 2014). With sufficient analytical foundations, the potentials of D2D communications are fully approved. However, achieving above engineering merits is not the primary design goal of D2D ProSe in

Release 12. Instead, a new feature of supporting communications with or without coordination from the network is created, which leads to two components in Release 12 D2D ProSe. (i) Through network-assisted discovery, UEs in close physical proximity are able to discover the existence of each other. (ii) Direct communications among two UEs are able to take place with or without supervision from the network (Wu et al. 2016). This feature saves scarce resources of the network for data exchanges when public safety events occur and empowers local communications among UEs outside network coverage. On the other hand, many public safety applications may require communications among a group of users. For example, users in a police team or a fire brigade team may share messages/voice with all other members in the team. For such a case, group communications are thus a technology to enable one-to-many communications.

To further extend the UE-to-network relay in Release 13, many countries adopting Release 12 as their public safety networks also decide to maintain PC5 interface unchanged (which will be elaborated later). In this case, only Layer 1 and Layer 2 of PC5 are insufficient to support the UE-to-Network relay, and additional procedures on the top of Layer 1 and Layer 2 (i.e., in Layer 3) should be needed to exploit Layer 1 and Layer 2. As a result, the UE-to-network relay in Release 13 is a Layer 3-based relay.

In Release 12, solely the D2D discovery for the in-coverage scenario and D2D broadcasting communications are solely supported. However, for public safety, an in-coverage UE may frequently move outward and toward eNB coverage, and therefore providing service continuity for D2D ProSe is of crucial importance. Fundamentally based on Release 12, the capabilities of discovery and communications in Release 13 are extended to support the partial-coverage and the out-of-coverage scenarios. Particularly for the transition from the in-coverage scenario to the partial-coverage scenario, an in-coverage UE is able to facilitate communications between an eNB and an out-of-coverage UE. For this

purpose, this feature is known as the Layer 3-based UE-to-network relay (Lien et al. 2016b).

Foundations

Since the major purposes of Release 12 D2D ProSe are to share the traffic load of the network when a large amount of data exchanges are needed in public safety events and to enable local voice/data communications when a part of or all eNBs are unavailable, three scenarios are supported by Release 12 and Release 13 D2D ProSe, as shown in Fig. 1:

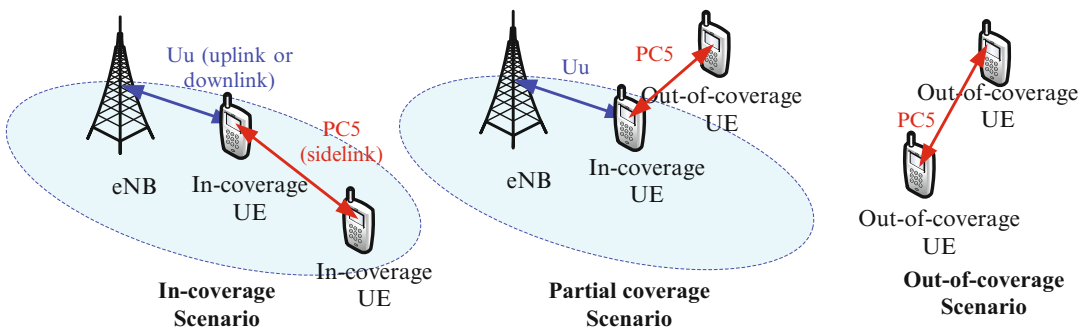
- **In-coverage** This scenario indicates that all the considered UEs are within eNB coverage (in-coverage UEs) to receive services/signals from an eNB.
- **Out-of-coverage** This scenario indicates that all the considered UEs are outside eNB coverage (out-of-coverage UEs) and cannot receive services/signals from an eNB.
- **Partial-coverage** This scenario indicates that some UEs are within eNB coverage, while other UEs are outside eNB coverage.

To support these scenarios, a new air interface known as PC5 is provided in addition to the conventional Uu interface. Different from Uu (defining the link between an eNB and a UE (to support downlink or uplink transmissions)), PC5 is the interface between two UEs (to support **sidelink** transmissions, in contrast to uplink and downlink), as shown in Fig. 1. In Release 12, PC5

is a connectionless interface. Unlike Uu in which a receiver is able to report messages back to a transmitter (these messages include channel state information, precoding matrix index, rank index, acknowledgment, or negative acknowledgment), feedback channels are not supported over PC5. In Layer 1, PC5 reuses the radio resources on the physical uplink shared channel (PUSCH) of Uu based on a twofold reason. Firstly, in the PUSCH, transmitters are UEs and the receiver is an eNB. When a UE is uploading data to an eNB, other UEs are able to perform D2D ProSe concurrently if the interference from a D2D link to the eNB is limited. However, in the physical downlink shared channel (PDSCH), a common transmitter is an eNB. In this case, if two UEs perform D2D communications concurrently along with a downlink transmission, the eNB becomes a strong interference source. Therefore, operating PC5 at the PUSCH reaches a higher spectrum efficiency. Secondly, the single-carrier frequency-division multiple access (SC-FDMA) is adopted for the PUSCH in LTE/LTE-A, while the orthogonal frequency-division multiple access (OFDMA) is adopted for the PDSCH. Considering that the peak-to-average power ratio (PAPR) in the SC-FDMA is smaller than that in the OFDMA, operating PC5 at the PUSCH also reaches a higher energy efficiency.

In PC5, there are two modes of resource allocations:

- **Mode 1** An eNB schedules radio resources for a UE to transmit direct data and control information for D2D communications.



Sidelink Transmissions and Proximity Services in LTE-A and Beyond, Fig. 1 Release 12 D2D ProSe support three scenarios: in-coverage, out-of-coverage, and partial-coverage

- **Mode 2** A UE on its own “randomly” selects resources from the pre-configured resource pools to transmit direct data and control information for D2D ProSe.

The Mode 1 resource allocation is supported by in-coverage UEs, while the Mode 2 resource allocation can be supported by both in-coverage and out-of-coverage UEs. If an in-coverage UE supports the Mode 2 resource allocation, the resource pools for this UE may be configured by eNBs. When a transmitter UE determines the radio resource occupation for D2D ProSe (either scheduled by the eNB or randomly selected by the UE), the transmitter UE needs to inform the receiver UE about its radio resource occupation. This information is referred to as sidelink control information (SCI). The radio resources used to transmit SCI also follow the Mode 1 and Mode 2 allocations. In other words, before data transmissions, a transmitter UE needs to decide the radio resources for transmitting SCI (either scheduled by the eNB or randomly selected by the UE). However, this radio resource occupation for SCI is unknown by a receiver UE. Therefore, a receiver UE needs to perform blind detection of radio resources carrying SCI.

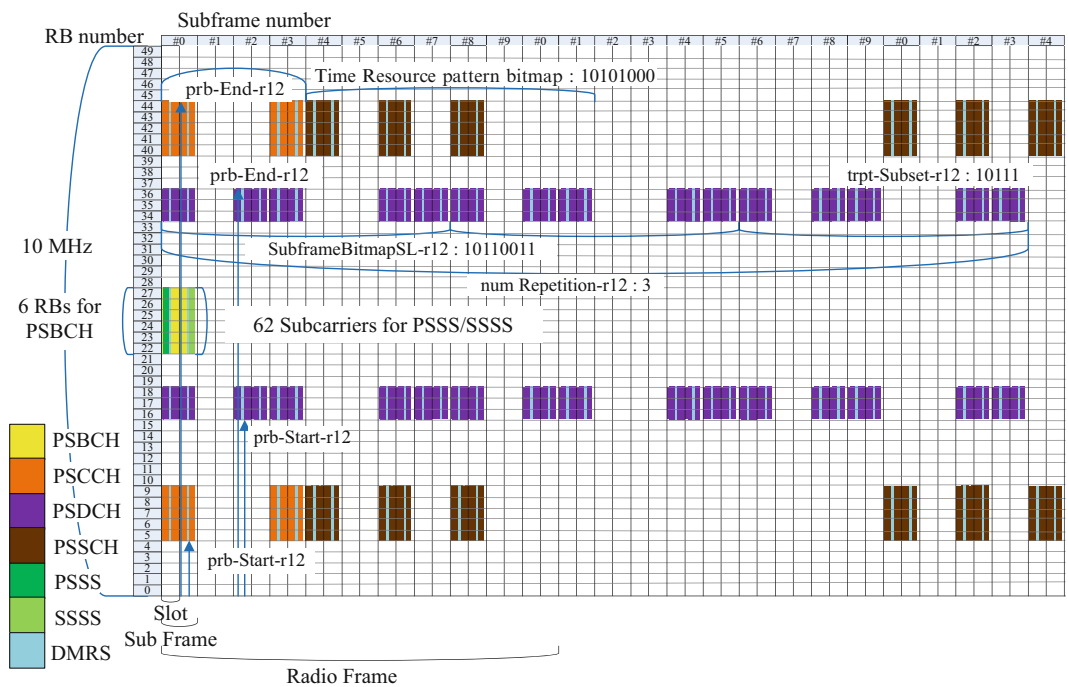
In PC5, the following sidelink physical signals and channels are supported, and the time-frequency resource locations of these sidelink physical signals/channels are illustrated in Fig. 2a:

- **Sidelink Synchronization Signals (SLSSs)** In LTE/LTE-A, synchronous operations are adopted. That is, all eNBs and all UEs should follow the same timing reference. To achieve this goal, in Uu, each eNB should broadcast the primary and secondary synchronization signals (carrying the timing reference) for all UEs to synchronize to. However, in the partial-coverage and out-of-coverage scenarios, there are UEs outside eNB coverage. Thus, only eNBs broadcasting synchronization signals are not enough. To enable all eNBs and all UEs (including in-coverage UEs and outside coverage UEs) following the same

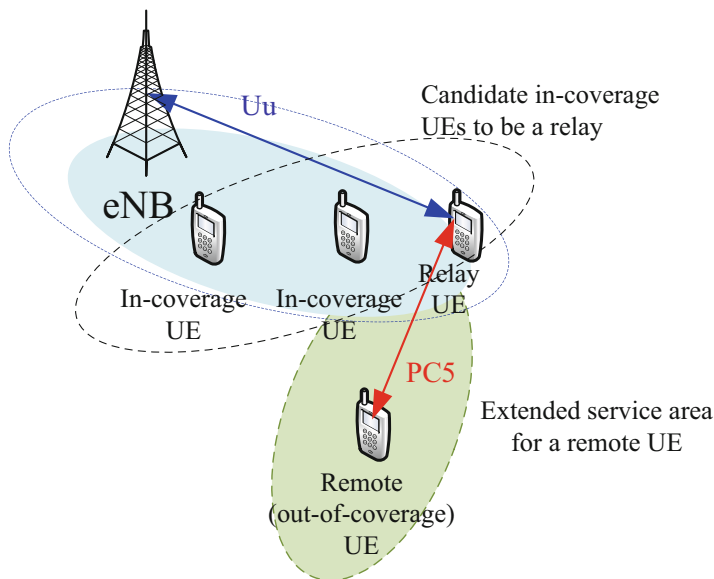
timing reference, in the partial-coverage scenario, an in-coverage UE can transmit its timing reference for outside coverage UEs to synchronize to. In the out-of-coverage scenario, an outside coverage UE is also able to transmit its timing reference for other outside coverage UEs to synchronize to. An eNB or a UE transmitting its timing reference is referred to as a **D2D synchronization source**. Therefore, there can be multiple collocated D2D synchronization sources. The transmitted signals carrying the timing reference are thus referred to as the SLSSs. By this definition, an eNB is certainly a D2D synchronization source, and the corresponding SLSSs are the Release 8 primary and secondary synchronization signals. In Release 12, there are two types of SLSSs: primary SLSS (PSSS) and secondary SLSS (SSSS).

- **Physical Sidelink Broadcast Channel (PSBCH)** In addition to the SLSSs, a D2D synchronization source also transmits the PSBCH, which carries information to support D2D synchronization, such as the frame number on the sidelink, the system bandwidth, the ID of the synchronization source, the type of synchronization source (i.e., the timing reference is derived by a UE or an eNB), the time division duplex (TDD) or frequency division duplex (FDD) configuration, and the stratum level (see Lien et al. (2016a) for the D2D synchronization procedure).
- **Discovery Signal and Physical Sidelink Discovery Channel (PSDCH)** The purpose of D2D discovery is to identify the existence of neighboring UEs. In Release 12, D2D discovery is only supported in the in-coverage scenario, and D2D discovery and D2D communications are two independent procedures. In other words, facilitating D2D communications is not the sole goal of D2D discovery, which also facilitates other public safety tasks. On the other hand, D2D communications are able to take place even without D2D discovery (without knowing the existence of neighboring UEs) through broadcasting. In Layer 1, two UEs discover each other via exchanging the discovery signals through the PSDCH

a



b



Sidelink Transmissions and Proximity Services in LTE-A and Beyond, Fig. 2 (a) Time-frequency resource locations of sidelink physical channels. (b) Layer 3-based UE-to-network relay in enhanced D2D ProSe

within a discovery period. During each discovery period, two types of D2D discovery are supposed:

- **Type 1:** Radio resources for discovery signal transmissions are allocated on a non-UE-specific basis. That is, the allocated

radio resources can be used by all UEs to transmit their discovery signals.

- **Type 2:** Radio resources for discovery signal transmissions are allocated on a per-UE-specific basis. That is, the allocated radio resources are only utilized by a specific UE to transmit the discovery signal:

Type 2A: Radio resources are allocated at each discovery signal transmission time instant.

Type 2B: Radio resources are semi-persistently allocated for discovery signal transmissions.

For the Type 1 discovery, the radio resources for transmitting the discovery signals can be shared by multiple UEs, and therefore this type of discovery can be supported by in-coverage UEs or UEs outside coverage. However, for the Type 2 discovery, the radio resources are allocated to each particular UE, and therefore this type of discovery is supported by in-coverage UEs, by which the radio resources are scheduled by an eNB. The PSDCH carries physical discovery signals which map the application layer discovery messages to physical signals for discovery (see a complete D2D discovery procedure from the application layers to Layer 1 in Lien et al. 2016a).

- **Physical Sidelink Control Channel (PSCCH) and Physical Sidelink Shared Channel (PSSCH)** Similar to the PUCCH and PUSCH used for uplink control and data transmissions, respectively, the PSCCH and PSSCH are used for Layer 1 control and data transmissions in D2D ProSe. The PSSCH carries data transmissions (e.g., voice) from upper layers via a unicast bearer or a multimedia broadcast multicast service (MBMS) bearer (as discussed later). On the PSCCH, a D2D transmitter can transmit SCI format 0, which carries the MCS used for D2D communications, the frequency-hopping flag to indicate whether to activate frequency hopping in D2D communications, the RB assignment and hopping resource allocation to indicate the RB allocation, T-RPT as aforementioned, and the timing

advance (TA) to indicate the timing difference between the downlink timing and the uplink timing. Please note that, in the conventional LTE/LTE-A Uu interface, eNBs broadcast synchronization signals in the downlink frequency band (for the FDD mode). By receiving these synchronization signals, each UE is able to synchronize to an eNB in the downlink frequency band. However, the timing reference of a UE does not align with the timing reference of an eNB in the uplink frequency band. To schedule an uplink transmission in uplink channels, an eNB thus should provide information of TA to inform a UE about the timing difference between the downlink frequency band and the uplink frequency band. Similarly, for D2D ProSe, an in-coverage UE should synchronize to an eNB in the downlink frequency band. However, sidelink channels reuse the frequency bands of uplink channels. Although SLSSs can be transmitted over the sidelink to facilitate synchronizations in the partial-coverage and out-of-coverage scenarios, TA to indicate the timing difference between the downlink channels and sidelink channels is still necessary, particularly in the partial-coverage scenario. Similar to the Uu interface that TA is carried by the control channel, TA in sidelink is also carried in the PSCCH.

Since sidelink signals/channels reuse the same radio resources of the PUSCH, it is possible that sidelink signals/channels utilize the radio resources those have been occupied by channels/signals on the PUSCH. In this case, the PUSCH has a higher priority to utilize the radio resources. It is also possible that multiple sidelink signals/channels may occupy the same radio resources, and the SLSSs/PSBCHs have the highest priority among other sidelink signals/channels.

In the partial-coverage and the out-of-coverage scenarios, out-of-coverage UEs may not receive control signals from an eNB. In this case, unlike the in-coverage scenario where radio resources for D2D discovery are

scheduled by an eNB (through the Type 2 discovery), radio resources of UEs outside coverage may not be fully controlled by an eNB. Consequently, the Type 1 discovery is primitively employed to the partial-coverage and the out-of-coverage scenarios. Similarly, unlike the in-coverage scenario where the timing reference for synchronization is provided by an eNB, UEs outside coverage may not receive the primary/secondary synchronization signals from an eNB. Losing synchronization, D2D discovery may not be successfully performed at an outside coverage UE. To facilitate synchronization for D2D discovery in the partial-coverage and the out-of-coverage scenarios, the following two behaviors to transmit SLSSs are supported in Release 13:

- **Behavior 1** A UE in each discovery period reuses the procedure of SLSS transmissions in Release 12.
- **Behavior 2** A UE in each discovery period (in which UE transmits discovery signals) transmits SLSSs every 40 ms.

In Release 12, the purpose of D2D discovery is solely for public safety, while in Release 13, the purposes of D2D discovery in the in-coverage scenario can be extended to nonpublic safety (e.g., commercial purposes). For an out-of-coverage UE transmitting the discovery signals using the Type 1 discovery, it should transmit the SLSSs following the Behavior 2. On the other hand, for an in-coverage UE transmitting the discovery signals using the Type 1 discovery, it should transmit the SLSSs following the Behavior 1 for a nonpublic safety purpose, and an eNB is able to configure to adopt the Behavior 1 or the Behavior 2 for public safety.

To initiate a relay discovery, an eNB should broadcast information associated with the relay discovery, to indicate that the relay discovery can be currently supported. In addition to the broadcast signaling, an eNB may also announce the transmission resource pool(s) and/or reception resource pool(s) for the relay discovery. Such a broadcast signaling may directly activate an in-coverage UE to be a relay

(UE) for the relay discovery, even though this in-coverage UE is in the idle mode. Through the broadcast signaling, an eNB may indicate the minimum and/or the maximum Uu link quality thresholds, by which a relay UE is able to autonomously start/stop the relay discovery. Nevertheless, if an in-coverage UE only receives the broadcast signaling from an eNB while the transmission resource pool(s) is not received, then an in-coverage UE is able to initiate a request for the relay discovery. Upon receiving such a request, an eNB may send the “dedicated signaling” to configure the UE (should be in the connected mode) to be a relay.

As aforementioned, due to mobility, an in-coverage UE may frequently move toward/outward coverage of an eNB. To provide the service continuity and extend the service range of an eNB, an in-coverage UE may be selected as a relay UE to relay data between an eNB and a remote UE, as shown in Fig. 2b. The procedure to select an appropriate relay UE is known as the relay selection. Due to fading channels and mobility, a selected relay UE may not always be suitable for the data relay all the time. When a relay UE is no longer suitable to support signal relay, a new relay UE may need to be selected, and this procedure is known as the relay reselection.

To select/reselect an appropriate relay UE, the link quality (e.g., reference signal receiving power, RSRP) both in PC5 and in Uu should be taken into consideration. Since a relay UE is generally in-coverage while a remote UE could be outside coverage, the link quality in PC5 may induce a larger impact than that in Uu to the performance in UE-to-network relay. To quickly adapt to the channel variation in PC5, a remote UE is able to trigger relay reselection.

Key Applications

Push-to-talk (PTT) services are major applications being supported by D2D ProSe, which provide a method for two or more users to engage in voice communications. Using PTT services, users request to transmit by means of

pressing a button on a UE. Therefore, each UE may autonomously initiate a voice transmission. The MCPTT is an enhancement of PTT services suitable for mission-critical scenarios based upon 3GPP evolved packet system (EPS) services. The MCPTT service is intended to support voice/audio communications among several users (i.e., group communications), where each user has the ability to gain access to the permission to talk in an autonomous manner. The normal operation in group communications is that one user in a group is allowed to talk at a time, to avoid interference in a group. The MCPTT thus further provides mechanisms to arbitrate among multiple transmission requests in contention. These mechanisms are also known as the floor control and priority control. In other words, when multiple requests occur, the determination of which user's request is accepted and which users' requests are rejected/queued is based upon the respective priorities in contention. For example, a team leader of a police group communicating with a sniper on the roof needs to give a "shoot" or "do not shoot" voice message in time. This request may be associated with a high priority.

Distinguishing priorities among transmitted traffic from a single user and from multiple users is of crucial importance for a public safety network. When a user presses the button to transmit an urgent voice message, a public safety network does not have to wait for transmission termination of currently ongoing traffic with a lower priority. Instead, a public safety network shall suspend the ongoing transmissions immediately and shall yield the radio resources to support the urgent message (with a higher priority). This operation is known as preemption or interruption. In D2D ProSe, the granularity to assign priorities to traffic from users is a packet (i.e., D2D ProSe support per packet priority, PPP). Since the Internet Protocol (IP) is supported by LTE-A, the definition of a packet in the priority control and floor control mechanisms is an IP packet.

In conventional mobile networks, a class of schemes to properly arrange transmissions of different traffic flows according to the corresponding priorities are known as floor control. These

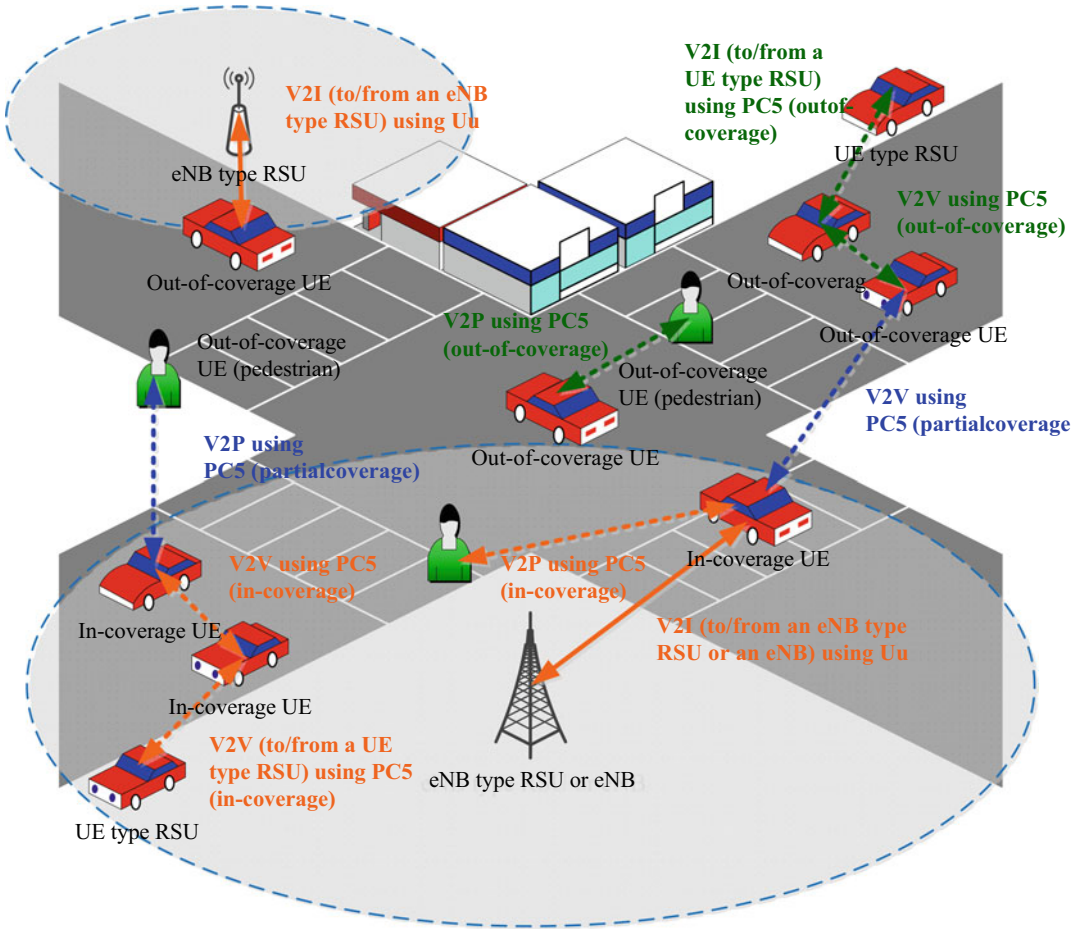
schemes include first-come-first-served (FCFS), token passing, round robin, etc. Floor control is also capable to limit the time of a user's talk (hold the floor), which thus permits requests of the same or lower priorities a chance to regain the floor. Floor control can be implemented in the application layer and therefore imposes no impact to the lower layer designs. However, as an MCPTT packet of a higher priority attempts to be transmitted, it is a challenge to interrupt the ongoing MCPTT transmission of a lower priority and release the radio resources immediately at lower layers. As a result, solely using floor control in the application layer, the latency performance to serve an urgent packet transmission may be a critical concern.

In lower layers, practicing transmission interruption is also a very challenging issue. For Mode 1 resource allocations, since an eNB schedules radio resources for all UEs, radio resource allocations for packets of different priorities can be fully controlled by an eNB. For Mode 2 resource allocations, each UE autonomously accesses radio resources. Due to the nature of a half-duplexing radio, when a UE is transmitting a low-priority packet, it may not receive a request of a high-priority packet transmission. As a result, a high-priority packet transmission still has to wait for a low-priority packet transmission.

To provide sufficient opportunities for a high-priority packet transmission particularly in Mode 2 resource allocations, Layer 2 of D2D ProSe supports a configurable mapping of PPP levels to different Mode 2 resource pools. In other words, there is a one-to-one mapping to associate each PPP level with a resource pool(s). As a result, when a high-priority packet needs to be transmitted, a resource pool(s) is reserved such that the packet can be transmitted immediately.

Future Directions

3GPP vehicle-to-everything (V2X) radio access aims at enabling a vehicle to exchange data with other vehicles (vehicle-to-vehicle, V2V), pedestrians (vehicle-to-pedestrians, V2P), and infras-



Sidelink Transmissions and Proximity Services in LTE-A and Beyond, Fig. 3 3GPP V2X includes V2V, V2P, and V2I. Through reusing PC5 interface design-

ated for direct UE communications, in-coverage, out-of-coverage, and partial-coverage scenarios can be supported

structure (vehicle-to-infrastructure, V2I), where infrastructure can be an eNB-type RSU or a user equipment (UE)-type RSU. The supported scenarios in V2X taxonomy are illustrated in Fig. 3. In V2I, if infrastructure is an eNB-type RSU, then Uu interface is reused. In V2V, V2P, and V2I (in the case that infrastructure is a UE-type RSU), PC5 is utilized.

On Uu, radio resources for uplink/downlink transmissions are allocated/scheduled by an eNB. However, for PC5, there are two radio resource allocation schemes (i.e., Mode 3 and Mode 4) depending on the following deployment scenarios. Please note that Mode 1 and Mode 2 are used for Release 12 and 13 D2D:

- **In-coverage** In this scenario, all vehicles/pedestrians are under coverage of an eNB, and both of the following two resource access modes can be adopted. (1) Radio resources used for sidelink transmissions are scheduled by the eNB, which is known as Mode 3. (2) Each vehicle/pedestrian on its own randomly selects radio resources from the pre-configured resource pool(s) to perform sidelink transmissions, which is known as Mode 4.
- **Out-of-coverage** In this scenario, all vehicles/pedestrians are out of eNB's coverage, and only Mode 4 is supported as there is no eNB able to perform resource scheduling.

- **Partial-coverage** In this scenario, a part of vehicles/pedestrians are in-coverage, while other parts of vehicles/pedestrians are out-of-coverage. Both Mode 3 and Mode 4 can be applied to this scenario.

Since there is no resource coordination among UEs in Mode 4, multiple transmitter UEs may select the same radio resource(s) as their PSCCHs. In this case, these PSCCHs collide with each other, and receiver UEs may lose the following PSSCHs. To mitigate PSCCH collisions, an enhancement of Mode 4 is provided in 3GPP Release 14 V2V. When a transmitter vehicle attempts to launch a transmission burst, it senses the resources in a measurement window and only selects resources not being occupied by other transmitters to transmit PSCCH. Nevertheless, as the resources carrying PSSCH are also randomly selected by each transmitter, it is still possible that there is no collision among PSCCHs, while collisions may occur among PSSCHs from different transmitters (Lien et al. 2017).

In Release 12 and 13 PC5, PSCCH and PSSCH are multiplexed in the time domain. However, to reduce radio access latency in V2V, PSCCH and PSSCH are multiplexed in the frequency domain. In this case, a receiver needs to receive signals over all resources within a particular time window (i.e., PSCCH/PSSCH transmission window), blindly detect corresponding PSCCH, and retrieve data at PSSCH. This design may not be cost-efficient in terms of complexity and energy, as the “space” for blind detection is largely extended. Nevertheless, complexity and energy may not be issues for a vehicle in practice. This design further enhances reliability in data reception, as a receiver receives signals within a PSCCH/PSSCH transmission window, rather than randomly selecting some resources to perform detection.

Cross-References

- [Energy Efficiency of Cellular Networks](#)
- [Relaying in LTE-Advanced](#)

References

- 3GPP TR 36.843 V12.0.0 (2014) Study on LTE device to device proximity services: radio aspects
- Ferrus R, Sallent O, Baldini G, Goratti L (2013) LTE: the technology driver for future public safety communications. *IEEE Commun Mag* 51(10): 154–161
- Lien SY, Chien CC, Tseng FM, Ho TC (2016a) 3GPP device-to-device communications for beyond 4G cellular networks. *IEEE Commun Mag* 54(3): 29–35
- Lien SY, Chien CC, Liu GST, Tsai HL, Li R, Wang YJ (2016b) Enhanced LTE device-to-device proximity services. *IEEE Commun Mag* 54(12): 174–182
- Lien SY, Deng DJ, Tsai HL, Lin YP, Chen KC (2017) Vehicular radio access to unlicensed spectrum. *IEEE Wirel Commun* 24(6):46–54
- Liu J, Kato N, Ma J, Kadowaki N (2015) Device-to-device communication in LTE-advanced networks: a survey. *IEEE Commun Surv Tutor* 17(4):1923–1940. Fourthquarter
- Peng M, Li Y, Quek TQS, Wang C (2014) Device-to-device underlaid cellular networks under Rician fading channels. *IEEE Trans Wirel Commun* 13(8):4247–4259
- Wu Z, Park VD, Li J (2016) Enabling device to device broadcast for LTE cellular networks. *IEEE J Sel Areas Commun* 34(1):58–70
- Yu CH, Doppler K, Ribeiro CB, Tirkkonen O (2011) Resource sharing optimization for device-to-device communication underlying cellular networks. *IEEE Trans Wirel Commun* 10(8):2752–2763

Signal Detection

- [Spectrum Sensing with Power Recognition](#)

Simulation of Diffusion-Based Molecular Communications

- [Modeling Approaches for Simulating Molecular Communications](#)

Situation Awareness in Smart Grids

Yawei Wang² and Chen Zhang^{1,2,3}

¹Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

²The George Washington University, Washington, DC, USA

³Southeast University, Nanjing, China

Synonyms

[Situational awareness](#)

Introduction

Situation awareness (SA), which was first clearly proposed by [Endsley](#) in (1988), has attracted the interest of practitioners and researchers in a variety of domains, including aviation, air traffic control, ship navigation, health care, military command and control operation, and environment ([Endsley 1995](#)). The early research on SA mainly took place in the aviation and military domains ([Parasuraman et al. 1992](#)). In order to conduct SA-related research and apply SA into different domains, several authors have proposed definitions of SA with the description of mechanisms developing a shared understanding of a situation. For instance, [Nofi \(2000\)](#) defined SA as a shared awareness of a particular situation, and [Perla et al. \(2000\)](#) claimed that shared SA implies that we all understand a given situation in the same way. [Rodney Wellens \(1993\)](#) described SA during collaborative activities through a model of distributed decision-making and defined SA as “the sharing of a common perspective between two or more individuals regarding current environmental events, their meaning and projected future.” [Endsley's \(1989\)](#) approach is different in that they defined SA as the degree to which every team member possesses the SA required for his or her responsibilities. With [Bass et al. \(1999\)](#) realizing the functional model of network situation

awareness based on multi-sensor and data fusion, the concept of SA is growingly introduced into cyberspace ([Lif et al. 2017](#)). In these fields, the SA is stated as a process of continuous extraction of systemic information along with integration of the information with previous knowledge to form coherent mental pictures and use of those pictures in directing future perception and anticipating future need. As the next generation of power supply system, smart grid is a complex environment with many interactions among human operators, data collectors, and human-machine interfaces. The composition and characteristics of smart grid make the development and maintenance of a sufficient SA a highly challenging task.

Historical Background

Although many attempts have been made to form a universally accepted definition of SA, it is still difficult to formulate a comprehensive definition that captures all aspects of SA. According to the situation awareness and safety survey proposed by [Stanton et al. in \(2001\)](#), three SA definitions including three-level model ([Endsley 1995](#)), the perceptual cycle model ([Smith and Hancock 1995](#)), and activity theory model ([Bedny and Meister 1999](#)) have been widely accepted and applied by the majority of circumstances within different domains.

Three-Level Model

[Endsley's](#) three-level model defined SA as the perception of the elements in an environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.

- **Level 1 SA – Perception:** As the lowest level of SA, system operators need to gradually receive the information related to their goals. For instance, in the context of aviation, SA is associated with the pilot's perception of information from aircraft instrumentation, weather condition, the behavior of the aircraft, and

the terrain control. Besides collecting the data individually, Bass et al. (1999) proposed a function model of SA based on multi-sensor and data fusion which demonstrated that the next-generation network's intrusion detection system (IDS) needs to fuse massive data collected by heterogeneously distributed network sensors so as to realize the cyberspace SA.

- **Level 2 SA – Comprehension:** The data collected from the Level 1 SA can be integrated and synthesized to produce an understanding of the relevance to present tasks. The comprehension level is a process of achieving situation understanding through identifying the most critical information of the SA elements from Level 1 SA. Using the perceived information, system operators are capable of detecting any deviations between the actual system state and the expected state and further decide on the most necessary and appropriate actions to implement. In contrast, a lower Level 2 SA would fail to proceed to an effective implementation of necessary actions which may mitigate the system capability and reliability.
- **Level 3 SA – Projection:** As the final step and the highest level of SA, projection is a prediction of the future behavior and status of system components and elements in the environments (e.g., projected potential aircraft conflicts). For example, anticipation of the predicted future situation provides air traffic controllers with sufficient time to prevent and resolve aircraft conflicts and plan a course of actions to achieve their goals. Accuracy of the projection is highly dependent upon the level of Level 1 SA and Level 2 SA. System operators with a high Level 3 SA are able to develop a set of strategies and responses to undesirable situations in advance.

Perceptual Cycle Model

In contrast with the Endsley's three-level model (Endsley 1995), Smith and Hancock's perceptual cycle model (PCM) (Smith and Hancock 1995) does not distinguish between the SA processes

and product. This approach is a development of Neisser's model (Neisser 1976) of the perceptual cycle which emphasized the impacts that the knowledge of how world works leads to a better understanding and interpretation to current situation and highlights complex environmental information modifies human thoughts which in turn directs their actions. Moreover, the interaction between the individual and the world will make the perceptual cycle model more dynamic than others.

Activity Theory Model

Activity theory model is an interactive subsystem approach to situation awareness that presented by Bedny and Meister in (1999). As an information processing approach, the model comprises orientational, executive, and evaluative eight functional blocks with dedicated purpose, connected through forward and feedback loops. The involvement of processes is dependent on the nature of the task and the goals of the individual. Each functional block in activity theory model holds a specific task in the development of situation awareness and structure of activity. The content of each functional block may vary based on the nature of dynamic situation.

SA Challenges in Smart Grids

Endsley and Connors (2008) described "SA demons," such as data overload, workload, and increased complexity, which are capable to every domain. In paper Endsley and Connors (2008) and Connors et al. (2007), these authors discussed the demons and challenges for SA in distribution control center and power transmission. In practice, it's very important to identify the factors governing SA in smart grids since electrical disturbance could be initiated by operator errors.

- *Software applications:* There are various applications and tools such as alarm processing, state estimator, and contingency analysis that are used by smart grid operators

to monitor and control the power system. Failures in any of these applications or tools can result in absence of early warning of deterioration of the system state, which may result in delayed and absent response from the operators and contributed to the large-scale blackout. The blackout in the US-Canada region in August 2003 (US Force 2003) is a typical example of failures in software applications which cause a huge economic loss for those two countries.

- *Measurements:* Real-time measurement is the key part of real-time grid analysis and sufficient SA maintenance. Operators use the data collected from simple sensors and advanced equipment such as the wide-area measurement systems (WAMS) and the phasor measurement units (PMU) to be aware of critical events. Missing real-time information, due to a failure in the communication systems or any measurement devices, may affect operator's SA and result in lower level of certainty and a false sense of smart grid security.
- *Environmental factors:* During emergencies, system operators often have to deal with a large amount of data and alarm overload. The critical information of the system may be neglected, and the attention of the operators may be attracted by a small part of data that is less important or even unrelated to the main objective if the data and the alarms are not effectively integrated and presented using advanced decision support systems. In addition, graphical user interface (GUI) is used to detect abnormal situations. However, the complexity in GUI may make the perception and interpretation of information collected a highly challenging task. Hence, besides using GUI, operators often have to scan several computer monitors and many pages of supervisory control and data acquisition (SCADA) data to reach a single decision.
- *Automation:* Increasing the level of automation can significantly reduce the operators' workload in smart grids. However, highly automated systems may leave operators detached from the control systems and even possibly become unaware of the actual system status. This is usually called "out-of-

loop" syndrome that was defined by Endsley (2018). In practice, operators may have a decreased ability to understand the system and to detect automation failures. As a result, the operator's SA is significantly compromised where human intervention is needed.

- *Individual factors:* Operators' experience, training, and professionalism are also the factors that may contribute to poor performance and insufficient SA in smart grids. For example, the union for the coordination of transmission of electricity (UCTE) incident (UCTE 2007) in 2006 is a representative example of the impact of individual factors. A lack of experience with insufficient training on the latest technologies or inappropriate actions may lead to cascading line outages and the large-scale blackout.
- *Data and information communication:* The Italian blackout (UCTE Investigation Committee et al. 2003) of 2003 illustrates the importance of effective data and information communication and coordination between system operators, especially in highly interconnected networks. Insufficient communication may lead to ineffective coordinated actions, which result in inadequate SA between system operators.

Key Applications

Situation awareness is involved in applications related to the development and maintenance of the smart grid reliability such as state estimation (Stae), anomaly detection (AnoD), and security analysis (SecA), which are deeply relied on the collected information from phasor measurement units (PMU) and wide-area measurement systems (WAMS) that are implemented in the smart grid.

- *SA integrated with StaE.* Research in CHEN et al. (2016) integrated the linear state estimator into situational awareness applications based on real-time synchrophasor data. Compared with the conventional state estimator,

the proposed system guarantees solutions and performs in real time at the PMU data rate. With guaranteed and accurate state estimation solution and advanced real-time data analytic and monitoring functionalities, the system is capable of assisting operators to assess and diagnose current system conditions for proactive and necessary corrective actions.

- *SA assisted with AnOD*. Motivated by growing concern about failures and attacks to distribution automation equipment, researchers employed micro-PMUs to proceed the anomaly detection (Jamei et al. 2017) and situation awareness in a distributed smart grid. By collecting the phasor data stream of a single micro-PMU without knowledge of the grid interconnection parameters and correlating the phasor streams across multiple micro-PMUs using electrical properties of the grid, two methods are employed to design the anomaly detection rules.
- *SA combined with SecA*. A lot of threats occur in a very short time but have huge impacts on smart grid and disturb its normal operation. To address this issue, Wu et al. (2016) proposed a security situational awareness mechanism for smart grid based on big data analysis. To get high accuracy for the big data analysis, fuzzy cluster-based association method is used to realize the preliminary analysis about the collected security facts from multiple devices over the long term. Then, game theory and reinforcement learning are introduced for security situational analysis.

Conclusion

This entry provides a fundamental of SA in smart grids including definitions, challenges in smart grid due to insufficient SA, and SA key applications in smart grids. Within these contexts, this entry provides a guideline for designing smart grid system in a way to achieve sufficient SA.

Cross-References

- [Cyber-Physical Security in Smart Grid Communications](#)

- [Smart Grid Communication Based on IEEE 2030 Standard](#)
- [Smart Metering, Specific Challenges, and Solution Approaches](#)

References

- Bass T et al (1999) Multisensor data fusion for next generation distributed intrusion detection systems. In: Proceedings of the IRIS national symposium on sensor and data fusion, vol 24. Citeseer, pp 24–27
- Bedny G, Meister D (1999) Theory of activity and situation awareness. *Int J Cogn Ergon* 3(1):63–72
- Chen H, Zhang L, Mo J, Martin KE (2016) Synchrophasor-based real-time state estimation and situational awareness system for power system operation. *J Mod Power Syst Clean Energy* 4(3):370–382
- Connors ES, Endsley MR, Jones L (2007) Situation awareness in the power transmission and distribution industry. In: Proceedings of the human factors and ergonomics society annual meeting, vol 51. SAGE Publications, Los Angeles, pp 215–219
- Endsley MR (1988) Design and evaluation for situation awareness enhancement. In: Proceedings of the human factors society annual meeting, vol 32. SAGE Publications, Los Angeles, pp 97–101
- Endsley MR (1989) Situation awareness in an advanced strategic mission (no. nor doc 89-32). Northrop Corporation, Hawthorne
- Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. *Hum Factors* 37(1):32–64
- Endsley MR (2018) Automation and situation awareness. In: Automation and human performance. Routledge, Boca Raton, pp 183–202
- Endsley MR, Connors ES (2008) Situation awareness: state of the art. In: 2008 IEEE power and energy society general meeting-conversion and delivery of electrical energy in the 21st century. IEEE, pp 1–4
- Jamei M, Scaglione A, Roberts C, Stewart E, Peisert S, McParland C, McEachern A (2017) Automated anomaly detection in distribution grids using upmu measurements. In: Proceedings of the 50th Hawaii international conference on system sciences, Jan 2017.
- Lif P, Granåsen M, Sommestad T (2017) Development and validation of technique to measure cyber situation awareness. In: 2017 international conference on cyber situational awareness, data analytics and assessment (Cyber SA), pp 1–8
- Niesser U (1976) Cognition and reality: Principles and implications of cognitive psychology. WH Freeman/Times Books/Henry Holt & Co
- Nofi AA (2000) Defining and measuring shared situational awareness. Technical report, Center For Naval Analyses, Alexandria
- Parasuraman R, Bahri T, Deaton JE, Morrison JG, Barnes M (1992) Theory and design of adaptive automation in aviation systems. Technical report, Cognitive Science Lab, Catholic University of America, Washington, DC

- Perla PP, Markowitz M, Nofi AA, Weuve C, Loughran J (2000) Gaming and shared situation awareness. Technical report, Center for Naval Analyses, Alexandria
- Rodney Wellens A (1993) Group situation awareness and distributed decision making: from military to civilian applications. In: *Individual and group decision making: current issues*, Lawrence Erlbaum Hillsdale, NJ, U.S.A, pp 267–291
- Smith K, Hancock PA (1995) Situation awareness is adaptive, externally directed consciousness. *Hum Factors* 37(1):137–148
- Stanton NA, Chambers PRG, Piggott J (2001) Situational awareness and safety. *Saf Sci* 39(3):189–204
- UCTE (2007) Final report system disturbance on 4 November 2006
- UCTE Investigation Committee et al (2003) Interim report of the investigation committee on the 28 September 2003 blackout in Italy. UCTE Report, 27 Oct 2003
- US Force (2004) Canada power system outage task. Final report on the August 14, 2003 blackout in the United States and Canada: causes and recommendations
- Wu J, Ota K, Dong M, Li J, Wang H (2016) Big data analysis-based security situational awareness for smart grid. *IEEE Trans Big Data* 4(99):1

Situational Awareness

► Situation Awareness in Smart Grids

60 GHz in WBAN and mm-Wave Technologies

Stefan Bruendl¹, Hua Fang^{2,3}, and Honggang Wang⁴

¹University of Massachusetts Dartmouth, North Dartmouth, MA, USA

²University of Massachusetts (UMass) Dartmouth, North Dartmouth, MA, USA

³UMass Medical School, Worcester, MA, USA

⁴Department of Electrical Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA

Synonyms

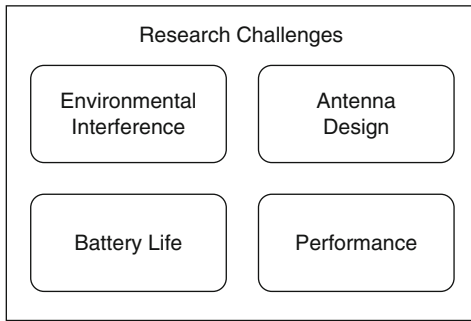
WBAN; Wireless Body Area Networks

Definition

A WBAN is a network that body sensors utilize to communicate with each other, by sending and storing data across sensors.

Background

Wireless body area networks (WBANs) are a highly popular communication system in the current market. As such it will transfer data with the help of a personal server and a Wi-Fi connection to a hospital so that doctors and caretakers can survey the results and statuses of the sensors on a patient's body. WBAN allows doctors to use special treatment sensors, such as EEGs and EKGs outside of the hospital, permitting users to leave the hospital and go about their daily lives without being required to stay in a hospital due to the equipment. WBANs also allow a network that hospitals can incorporate more easily into their patient survey programs so that doctors and nurses can react more easily to unusual readings of monitoring devices. Ever since initial proposals have been made for WBANs, the network has been getting stronger and more efficient. Currently WBANs operate on the 2.4 GHz frequency which is becoming increasingly overpopulated with modern technologies such as phones using Bluetooth and Wi-Fi, for example. One very promising solution to the problem of exhaustion of resources is the switch from a 2.4 GHz waveband to a higher range such as 5–10 GHz and, as will be discussed in this section, the 60 GHz waveband. Since WBANs are becoming so increasingly popular and since resources in the 2.4 GHz range are as limited as they are, switching to a larger waveband that can sustain more WBAN sensors seems like the optimal solution. As will be discussed in the next section, 60 GHz promises a lot of benefits and improvements over the 2.4 GHz waveband, but also brings with it some imperfections that have to be elaborated on and explored. There are still some issues that need attention, such as battery life, antenna design, maximum distance of transmission and performance, security issues, and



60 GHz in WBAN and mm-Wave Technologies, Fig. 1
The overarching challenges of 60 GHz WBAN research

environmental interference. These are the challenges researchers tackle when trying to optimize WBANs at 60 GHz with the major ones being shown in Fig. 1.

Potential Solutions for 60 GHz WBANs

Comparing 60 GHz to 2.4 GHz Waveband

The current standard waveband for almost all wireless applications and accessories is the 2.4 GHz waveband. It is used for all kinds of things, ranging from phones to laptops to Bluetooth and Wi-Fi signals, and has been this way since the 1950s. As our technology keeps improving alongside networks and wireless devices, the 2.4 GHz waveband becomes more and more cluttered with possibly interfering signals and may run out of usable frequencies in coming years, causing trouble across networks through incomplete signals and interruptions.

There have been some studies researching and comparing 2.4–60 GHz to see whether or not it would be beneficial to swap frequencies. One very crucial aspect of comparison is the carrier-to-interference ratio (CIR) of both wavebands. CIR compares the amount of successfully transmitted data through a network to the loss of information across a network to result in a ratio that can be used to model efficiency in transmission of data across nodes in the case of WBAN. CIR is most notable if other signals on the same wavelength are present. Especially in

hospitals, this then plays a large role considering the amount of medical equipment and the amount of WBAN sensors in a possibly densely populated area. A study considering movement of nodes within a system by moving body parts and nodes of a WBAN system attached to them shows that 60 GHz promises a stronger CIR compared to 2.4 GHz under the circumstances given (Wu et al. 2015).

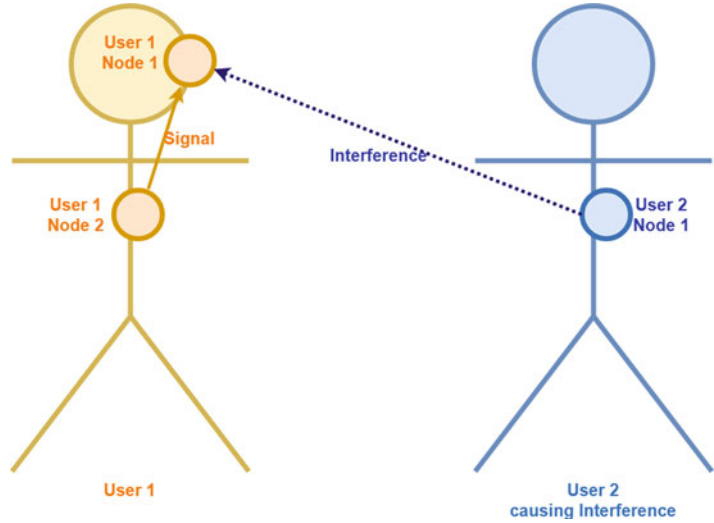
Going away from intra-network interference and taking into account inter-network interference (Fig. 2) show that CIR of 2.4 GHz is heavily affected by even one other BAN or WBAN network in the same area. 60 GHz becomes a lot weaker across larger distances which is why it can be beneficial to apply this type of wavelength rather than the 2.4 GHz frequency as WBAN systems increase. This ultimately allows densely populated areas such as hospitals to deploy multiple WBAN systems at once while losing the least amount of data as possible (Cotton et al. 2010).

Environmental Interferences

While the 60 GHz frequency doesn't struggle all too much with inter-network interference, it is a little undermined in its range. Only because it doesn't struggle much with other networks doesn't mean that it doesn't struggle with keeping its communication across nodes to a maximum. Environmental effects and humans affect the strength of the 60 GHz signal by blocking its line of sight or otherwise disturbing the frequency. A 60 GHz WBAN node with any of its antennas has rather poor penetration power and may lose some information across large distances or through obstacles and interfering objects in its way. This is nothing new; 2.4 GHz and other frequencies experience the same type of interference. It has large precision across short distances, but that fact diminishes if nodes are further apart since 60 GHz has large oxygen absorption loss and a poor value of penetration across objects. This is a problem that 60 GHz doesn't necessarily fix, but can hold up, compare to, and possibly improve upon previous methods such as the 2.4 GHz and ultra-wideband frequencies (Liu et al. 2015).

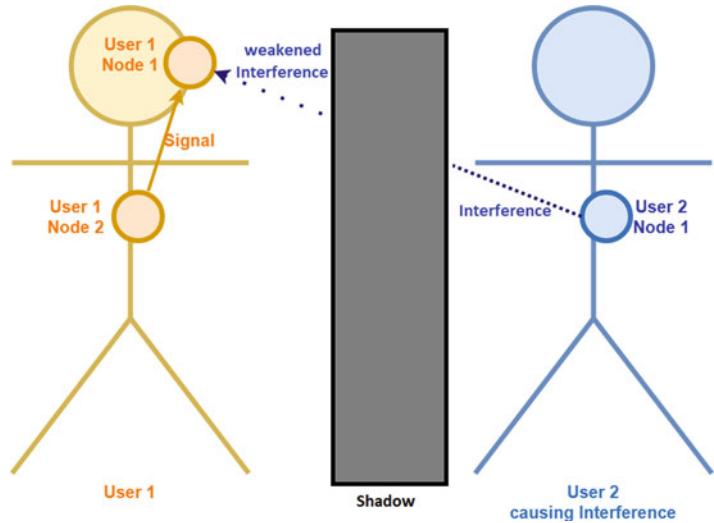
60 GHz in WBAN and mm-Wave Technologies, Fig. 2

Diagram demonstrating inter-WBAN interference



60 GHz in WBAN and mm-Wave Technologies, Fig. 3

Interfering signals are weakened by intercepting objects



A similar study found similar results by placing interfering signals (a blocking human system) between two systems of 60 GHz WBAN networks. Using the Friis formula, the results show that 60 GHz wavebands are interfered less if there is a blocking shadow signal to reduce the interference. The model the authors use for their research incorporates free space loss and the three users (interfered/interfering/blocking) to reach their conclusion: in the case of NLOS (non-line of sight), the interference strength of 60 GHz inter-WBAN is below power sensitivity (Akimoto et al. 2017). This means that even though 60 GHz waveband signals easily lose

strength and connectivity across larger distances, it can turn out to be a benefit rather than a drawback. The system encapsulates itself, only being able to communicate with nearby nodes and can't be interfered as easily by outside sources, especially if that outside source is intercepted by a blocking object such as a shadow in Fig. 3.

Researchers also have evaluated intra-WBAN interference alongside inter-WBAN interference to see if there may be any problems in short distance communication across WBANs. This was done by using a flashlight as a directional millimeter-wave WBAN to prove their study. It has been concluded that a millimeter-wave

WBAN does not consider body attenuation when using directional antennas, at least as described in specific example cases of intra-WBAN interference. This finding is very useful for future research as it shows that body attenuation is not a problem for line of sight (LOS) WBAN networks at 60 GHz. The intra-WBAN interference is almost negligible in this type of network and makes 60 GHz wavebands stand out from others (Akimoto et al. 2018).

In review, CIR is reduced since it is more easily intercepted than 2.4 GHz wavebands, there are less networks currently out in the market using 60 GHz as a waveband, and connection between nodes is stronger than other frequencies; 60 GHz seems to be the strongest contender that can currently be implemented in the medical field and as a communication tool.

Antennas Used in WBANs

As mentioned in the previous section, 60 GHz operates the strongest if used at low distances and has diminishing returns at larger distances. There are studies going on to develop methods and structures for antennas in a WBAN to stabilize connections and to maximize gain. There are many factors to consider when designing an antenna, such as detuning, changes in input impedance, changes in transmission power and accuracy, changes of radiation patterns, battery life, robustness and stability on the body, weight, and overall performance of the antenna over time. A crucial factor that will also need to be considered is the interfering obstacle that is the body of the user. Path loss on a human body ranges from 57 to 88 dB and can cause large amounts of lost packages in WBAN communication and has to be considered (Pellegrini et al. 2013).

Antennas that were considered by this source all have their benefits and drawbacks (Pellegrini et al. 2013). The Yagi-Uda antenna establishes a good connection, but doesn't optimize gain. An antenna using a microstrip-fed patch array is close to insensitive to phantoms and objects. Textile antennas have the benefit of being woven or otherwise inserted into clothes and that the human body has barely any effect on the antenna and the transmission of its signals.

Woodpile EBG-based antennas have a higher maximum directivity of +5 dB if the standard is comprised of monopole antennas. The final antenna described is the reconfigurable antenna that switches the beam direction of the radiation pattern which leads to better security, overcomes shadowing, minimizes unwanted illumination of body area, but otherwise doesn't see much use in WBAN systems (Pellegrini et al. 2013). Considering which antenna to use may depend on the desired structure of the WBAN on humans, be it that the nodes are attached to feet and arms or head and chest. Some benefit more than others in different situations.

While that article reviews the antennas that existed around 5 years ago, there are also advancements in design of antennas. Previous antennas in WBAN (prior to 2015) were already in a dual-band or broadband format but required additional components and structures to work (Zhu et al. 2016). In WBAN, minimizing effect on human bodies, minimizing the size of an antenna, and maximizing throughput are at the core of design of antennas. The proposed antenna attempts to improve upon its predecessors in order to improve performance. This has been happening for a long time (prior to 2010) and is an ongoing field of research. Designing an antenna is not limited to the factors mentioned above, however. While the efficiency of signals and attenuation along with size are large factors of design, optimization doesn't stop there.

Algorithms to Extend Battery Life

WBANs and their antennas, just like any other electrical device, require a power source. With WBANs, it can become tricky to implement batteries and energy-efficient methods of dealing with energy output as there are multiple ways to attach them to humans. Some WBANs may simply be attached to clothing, but some are implanted within the tissue and skin of a human. Once a battery runs out of power, it is extremely difficult to replace once within a patient's skin. Researchers have been looking for methods to reduce energy consumption and optimize battery usage to make this issue less apparent.

A proposed solution is to put the nodes in a WBAN to sleep if they are not in use. This method needs to take into consideration the types of transmission within a node. Messages can be delivered in parts and rather slowly or all at once. There are also different ways of assigning priority and emergency messages/packages to be sent. There have already been studies to observe and analyze these different types of packages and their consumption and that propose the above-mentioned solution through their findings (Jacob and Jacob 2017).

Combining power-saving strategies is also becoming more important. A very recent proposal (2018) is to use an optimal energy consumption algorithm based on an artificial bee colony structure (Yan et al. 2018). Nodes using this algorithm will be considering the remaining energy of a node, priority of nodes and messages, path costs of transmitting packages, and difference in path energy ratios when transmitting data. This combines previous studies under an algorithm that is based off a bee colony structure (Fig. 4). Packages find the most efficient and fastest path to travel across multiple or few nodes (depending on algorithm results) in order to maximize battery efficiency while keeping traffic of packages consistent. This algorithm is claimed to be better suited for WBAN systems than ant colony and genetic algorithms. While

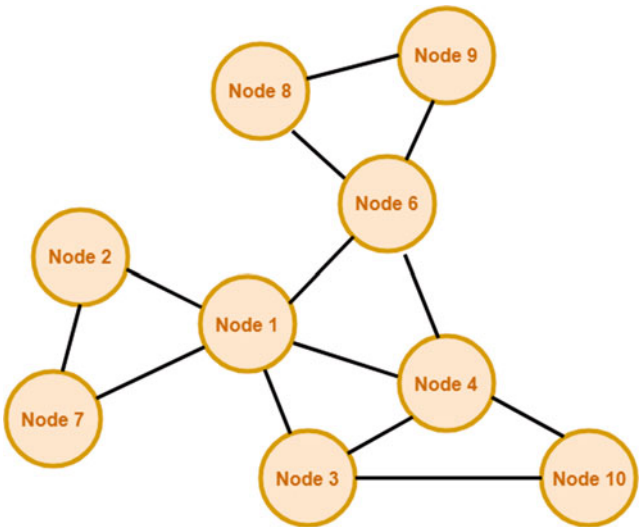
this algorithm maximizes battery efficiency, there are other methods and algorithms that specify in different aspects of WBAN systems, algorithms that improve nodes in different ways.

Algorithms to Enhance Performance

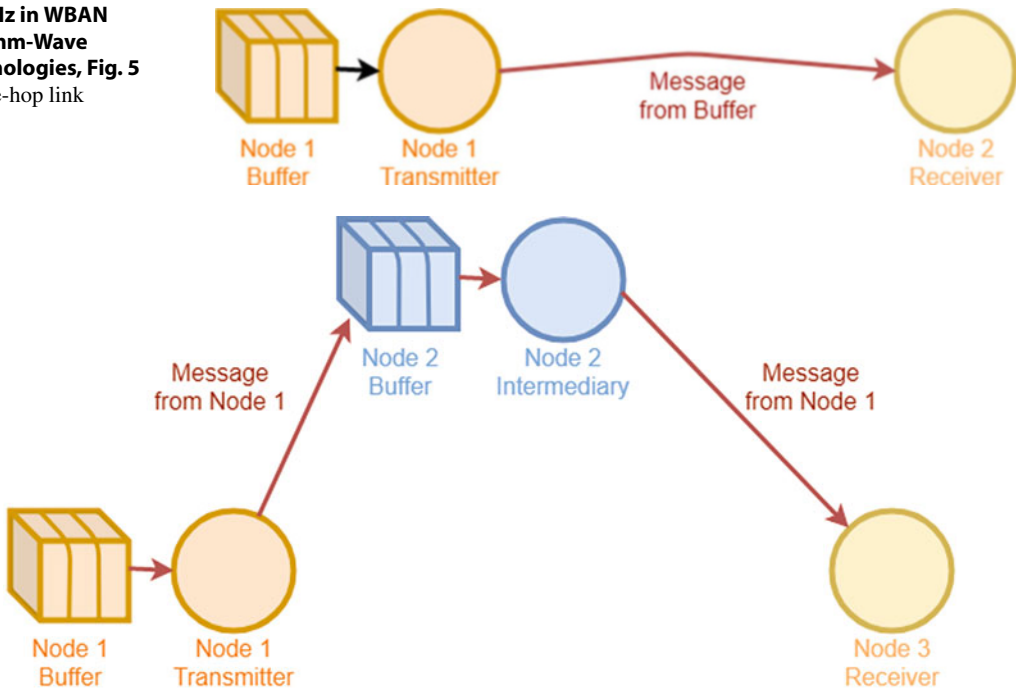
The previously mentioned artificial bee colony algorithm already considers communication across nodes, but mainly focuses on that one algorithm for transferring packages. There are however many multiple ways of getting packages from point A to point B especially if there are intermediary nodes that can be used to describe a different path. These ways include, but are not limited to, two-way relay cooperation links, one-way relay cooperation links, and direct transmission between WBAN sensors. Most systems specialize in one of these links and always use the same type of transmission for every type of package.

Studies to analyze and compare the performance of these transmissions have been made and come to a conclusion that single-hop links (Fig. 5), which work in a linear fashion to send information across nodes in a 1-to-1 relationship with a transmitter including a message buffer and a receiver are more energy efficient than dual-hop links (Fig. 6), which uses intermediary nodes as transmitters in the case of NLOS situations

60 GHz in WBAN and mm-Wave Technologies, Fig. 4 An example of an artificial bee colony node setup



60 GHz in WBAN and mm-Wave Technologies, Fig. 5
Single-hop link



60 GHz in WBAN and mm-Wave Technologies, Fig. 6 Dual-hop link

between two nodes. A solution proposed on the basis of this finding would use methods to also harvest energy from the environment while utilizing a single-hop transmission system (Mosavat-Jahromi et al. 2017).

A proposed communication system uses a hybrid switching scheme to quickly switch between these aforementioned methods of communication to maximize efficiency and speed while attempting to keep battery usage low (Waheed et al. 2018). This improves performance significantly for both LOS and NLOS since different links optimize one of either signal and since they can be switched to within short amount of times. This switching also results in a better packet-error-rate across nodes. Another benefit of this is that WBAN systems could be implemented with less implants and wearables due to a more effective communication framework.

Some ways of communication benefit under certain circumstances, and being able to switch between them and having an algorithm coordinate that switching is certainly a step in the right direction.

Simulation Platforms

While research is expanding in multiple directions, there is not a lot of hardware currently on the market to support testing at 60 GHz wave bands. There are some open-source and build-your-own type models that can be built to work with the frequency, but there are also some alternatives that may not be entirely accurate but useful to research as they can still arrive at general conclusions.

In general, the usage of a universal software radio peripheral (USRP) is recommended since it is standardized hardware for wireless communication. While models of USRP are not able to communicate at 60 GHz frequencies, they do have the ability to communicate beyond the 2.4 GHz threshold. The ability to do so allows USRPs to access different wavebands and therefore simulate and give similar results as if the simulation were done with a 60 GHz device. It is an alternative to test new research and methods, and while it isn't completely accurate, it is a step in the right direction, and results can be derived

for 60 GHz frequencies by looking at the results given by USRP.

One of the first accessible, build-your-own models for 60 GHz communication is a software-designed radio (SDR)-based, open-source front-end testbed published in early 2015. There were earlier machines that proposed to do the same, but this one is made with researchers and budgets in mind, making it easy to obtain all required parts and using the provided software. Like many other SDR-based machines, this device is used alongside a USRP. The combination of ease of construction and provided software made this model very valuable when it was released, but at this point in time, it is no longer the only one (Zetterberg and Fardi 2015).

A platform that was released at the end of 2016 called OpenMili (Zhang et al. 2016) expands on the functions of the previously mentioned testbed by allowing more customizability for users. This is a continued iteration of previous models and breaks through some very important limitations. The publication once again mentions in detail how all hardware components, which are easily obtainable, work together to provide a strong tool for 60 GHz experimentation. The key point that makes this model stick out is its customizability in a lot of different areas. The phased-array antenna included in the testbed, for example, is programmable and compatible with different interfaces and antennas. OpenMili is also optimized toward giga-samples per second (Gsps) sampling and Gsps processing rates to simulate a faster communication environment for 60 GHz communication. OpenMili declines the usage of USRP for communication as it seems to lower the sampling rate which is something OpenMili aims to optimize. It strays away from conventional models to create its own optimized communication network based off hybrid schemes, but not on USRP directly. As such it is able to perform at better speeds than previous testbeds at the time (Zhang et al. 2016).

Since then there has been a more recent, more promising device that tackles the lack of hardware to support 60 GHz communication. The testbed is called X60 and behaves similarly to the OpenMili platform. The SDR-based, multi-

gigabit WLAN, X60 60 GHz testbed has high customizability as well with adjustable MAC, PHY, and network layers alongside an antenna array that can be modified similar to OpenMili. The fact that X60 allows users to fully control MAC and PHY layers grants them the chance to test their algorithms and techniques on multiple layers while being able to also survey each layer fully. These additional features grant testers a greater potential in testing 60 GHz communication and performance of the X60 under different circumstances, including LOS and NLOS environments (Saha et al. 2017).

While the abovementioned devices are really promising for research in 60 GHz, their production cost is still rather high. There are not many 60 GHz testbeds out on the market as of right now, but the USRP-only alternative seems viable for low-scale testing and familiarizing with the concept of wireless communication beyond the current 2.4 GHz norm. Other testbeds are still on the rise, and some may soon even overshadow those mentioned here, but for now, there are not a lot of ways to test research done on 60 GHz other than with some testbeds.

Further Direction and Discussion

In all of these research challenges, a basic understanding of wireless communication is recommended to start delving into more detail about the chosen challenge. Areas such as environmental interference, inter-WBAN interference, and intra-WBAN interference will require to be able to correctly apply communication methods such as single- and dual-hop links or otherwise research new networking methods to improve current CIR and other issues that still arise during communication. Battery life and performance enhancing algorithms also go hand in hand. Optimizing one requires an understanding of the effects of the other and how to handle both. The challenge becomes interdisciplinary; between software and circuit design for antennas and handling communication strength and accuracy, WBANs at 60 GHz require strong design decisions and a result of application of all skills mentioned above.

Research in this field is looking for improvements in each field mentioned above, even hardware development. If funding is scarce, researchers should look to an alternative hardware to test new ideas on, such as USRP communication devices. With more funding and more knowledge, research can then expand into machines that accurately simulate 60 GHz wavelengths. Regardless of the challenge or field of 60 GHz research, there are a lot of resources to aid in understanding and learning to overcome these issues. Researching a field of challenge can aid in understanding the ins and outs, but hands-on experience will be very important in fully bridging the gap between theory and execution.

Conclusion

The 60 GHz waveband promises a bright future for WBAN networks. With reduced CIR, improved communication across short distances, difficult interception by other networks, and possibly better speeds and implementation among other points, 60 GHz improves what 2.4 GHz, UWB, and other frequencies had to offer. Ongoing studies in both 60 GHz and other frequencies may be applied to the same category of WBAN, with algorithms to solve remaining issues such as battery life and optimal antenna design. Switching systems from 2.4 to 60 GHz may take a while, but starting with WBAN, which is easy to switch over, it may happen eventually. It will have to; since we are running out of space in the 2.4 GHz waveband and since 60 GHz seems so promising, there doesn't seem to be a point not to. There have been vast improvements for optimizing path loss and communication speed and accuracy while also keeping battery life in mind, but those aren't perfect yet. Studies are still somewhat spread out, and, since there is no standard for 60 GHz WBAN systems yet, it is hard to combine all studies into one general design while implementing all improvements. Once these different fields of research reach a common optimized product, improvements can keep happening, but right now all information is spread out across disciplines and split between

2.4 GHz advancement and 60 GHz advancement. Both may or may not be applicable in either situation, but a standard should still be created to apply new solutions to.

Cross-References

- ▶ [Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks](#)
- ▶ [Energy Harvesting Technologies in Wireless Sensor Networks](#)
- ▶ [Molecular Communication for Wireless Body Area Networks](#)
- ▶ [Wireless Fading Channel Model for 5G and Future](#)
- ▶ [5G Wireless](#)

Acknowledgment This research is partly supported by the REU component of NSF 1744272 to Dr. Fang.

References

- Akimoto K, Kameda S, Motoyoshi M, Suematsu N (2017) Measurement of human body blocking at 60 GHz for inter-network interference of mmWave WBAN. In: 2017 IEEE Asia Pacific microwave conference (APMC), Kuala Lumpur, pp 472–475
- Akimoto K, Motoyoshi M, Kameda S, Suematsu N (2018) Measurement of on-body propagation loss for directional millimeter-wave WBAN. In: 2018 11th Global Symposium on Millimeter Waves (GSMM), Boulder, CO, USA, pp 1–3
- Cotton SL, Scanlon WG, Hall PS (2010) A simulated study of co-channel inter-BAN interference at 2.45 GHz and 60 GHz. In: The 3rd European wireless technology conference, Paris, pp 61–64
- Jacob A, Jacob L (2017) A green media access method for IEEE 802.15.6 wireless body area network. *J Med Syst* 41:179
- Liu X, Zhang H, Gulliver TA, Xu L (2015) Ranging performance with 60 GHz signals in indoor residential environments. In: 2015 IEEE Pacific Rim conference on communications, computers and signal processing (PACRIM), Victoria, pp 70–73
- Mosavat-Jahromi H, Maham B, Tsiftsis TA (May 2017) Maximizing spectral efficiency for energy harvesting-aware WBAN. *IEEE J Biomed Health Inform* 21(3):732–742
- Pellegrini A et al (2013) Antennas and propagation for body-centric wireless communications at millimeter-wave frequencies: a review [wireless corner]. *IEEE Antennas Propag Mag* 55(4):262–287

- Saha SK, Ghasempour Y, Haider MK, Siddiqui T, De Melo P, Somanchi N, Zakrajsek L, Singh A, Torres O, Uvaydov D, Jornet JM, Knightly E, Koutsonikolas D, Pados D, Sun Z (2017) X60: a programmable testbed for wideband 60 GHz WLANs with phased arrays. In: Proceedings of the 11th workshop on wireless network testbeds, experimental evaluation & characterization (WiNTECH'17). ACM, New York, pp 75–82
- Waheed M, Ahmad R, Ahmed W, Drieberg M, Alam MM (2018) Towards efficient wireless body area network using two-way relay cooperation. *Sensors* 18
- Wu X, Nechayev YI, Constantinou CC, Hall PS (2015) Interuser interference in adjacent wireless body area networks. *IEEE Trans Antennas Propag* 63(10):4496–4504
- Yan J, Peng Y, Shen D, Yan X, Deng Q (2018) An artificial Bee Colony-based green routing mechanism in WBANs for sensor-based E-healthcare systems. *Sensors (Basel, Switzerland)* 18(10):3268. <https://doi.org/10.3390/s18103268>
- Zetterberg P, Fardi R (2015) Open source SDR frontend and measurements for 60-GHz wireless experimentation. *IEEE Access* 3:445–456
- Zhang J, Zhang X, Kulkarni P, Ramanathan P (2016) OpenMili: a 60 GHz software radio platform with a reconfigurable phased-array antenna. In: Proceedings of the 22nd annual international conference on mobile computing and networking (MobiCom'16). ACM, New York, pp 162–175
- Zhu X, Guo Y, Wu W (2016) A compact dual-band antenna for wireless body-area network applications. *IEEE Antennas Wirel Propag Lett* 15:98–101

Sleep Monitoring Using WiFi Signal

Xuyu Wang, Chao Yang, and Shiwen Mao
Department of Electrical and Computer
Engineering, Auburn University, Auburn, AL,
USA

Synonyms

Remote health sensing using WiFi signal; Vital sign monitoring using WiFi signal

Definition

Sleep monitoring is to detect and analyze the activities of a subject during sleep, such as res-

piration rate, heart rate, and posture, to provide important information on the subject's health conditions (Pantelopoulos and Bourbakis 2010; Chen et al. 2013; Wang et al. 2018a,b). In addition, sleep quality can also be inferred by analyzing such vital signs. For example, monitoring the breathing signal while the patient is sleeping can help to detect sleep disorders, as well as sudden infant death syndrome (SIDS) (Hunt and Hauck 2006) and apnea (Nandakumar et al. 2015). However, most traditional sensors are intrusive, such as a capnometer (Mogue and Rantala 1988) or a pulse oximeter (Shariati and Zahedi 2005). A patient can hardly sleep comfortably with these sensors attached to the body. Because of the restrictions of traditional sensors, effective and comfortable sensors are in a great demand for long-term and contact-free vital sign monitoring.

Historical Background

Recently, sleep monitoring are implemented using different techniques, including RF radar, RFID, motion sensors, and acoustic-based sensors. Vital-Radio is an RF radar-based technique, which uses a frequency-modulated continuous wave (FMCW) radar to monitor breathing and heart rates (Adib et al. 2015). It requires a customized hardware using a large bandwidth from 5.46 to 7.25 GHz. Moreover, other radar-based vital sign monitoring techniques are proposed, including ultra-wideband radar (Salmi and Molisch 2011) and Doppler radar (Droitcour et al. 2009; Nguyen et al. 2016), which all require customized hardware with a relatively high cost and operates with a large spectrum band. RFID tags are used for vital sign monitoring too. For example, TagBreathe estimates breathing rates with phase information by grouping the signals with the same channel index (Hou et al. 2017). In fact, the system does not work very well in the US, where the RFID frequency range is from 902.5 to 927.5 MHz with 50 channels; the channel hopping among the 50 different frequencies leads to a large latency. To address this issue, the AutoTag system incorporates a recurrent variational autoencoder

for unsupervised apnea detection with RFID tags, as well as a frequency hopping offset mitigating technique to enable FCC-compliant RFID systems for realtime applications (Yang et al. 2018). Zephyr is a motion sensor-based monitoring technique, which can monitor breathing rates by fusing the data from the accelerometer and gyroscope available in many smartphones (Aly and Youssef 2016). In addition, HeartSense can even achieve a 1.03 bpm median absolute error for heart rate estimation using different gyroscope axes with a probabilistic model (Mohamed and Youssef 2017). The sleep monitoring system uses the built-in accelerometer to estimate breathing rate and body position with a smartwatch (Sun et al. 2017). For acoustic signal-based sensors monitoring, the apnea detection system employs an active FMCW sonar built in a smartphone to monitor the breathing signal (Nandakumar et al. 2015), while SonarBeat considers sonar phase information for breathing rate estimation with a continuous-wave (CW) radar (Wang et al. 2017a).

Although different techniques have been proposed for sleep monitoring, contact-free and long-term vital signs monitoring at low cost are still of high interest. Currently, WiFi-based systems have become an effective solution for sleep monitoring, as shown in Fig. 1, because of the wide availability and the low hardware

requirements of WiFi. Received signal strength (RSS) in WiFi networks has been used for vital sign monitoring in Abdelnasser et al. (2015). Moreover, by slightly modifying the device driver, channel state information (CSI) can be extracted from off-the-shelf WiFi network interface cards (NIC) (Halperin et al. 2010), which reflects represents fine-grained channel information. CSI amplitude (Liu et al. 2015) and phase difference (Wang et al. 2017b,c) have been leveraged for sleep monitoring. WiFi-based techniques for sleep monitoring will be reviewed in detail in the remainder of this article.

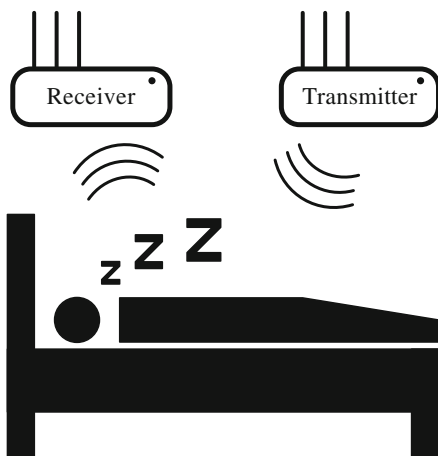
Preliminaries of WiFi Signals

WiFi standards, such as IEEE 802.11a/g/a/ac, all use the orthogonal frequency-division multiplexing (OFDM) technique in the physical layer (PHY), which can effectively overcome frequency selective fading in indoor environments. With OFDM, the total spectrum is divided into multiple orthogonal subcarriers, while data is encoded and mapped to these subcarriers to be transmitted using the same modulation and coding scheme (MCS) (Wang et al. 2016a, 2017d). At the receiver, the wireless signal is down-converted to the baseband. Using the serial-to-parallel (S/P) transform and fast Fourier transform (FFT), the subcarriers can be converted from the time domain to the frequency domain as in Fig. 2. For some commodity NICs, such as the Atheros AR9390 chipset (Xie et al. 2015) and Intel 5300 NIC (Halperin et al. 2010), there are open-source device drivers that allow reading the CSI for received packets, which represents fine-grained PHY channel measurements, containing rich information on channel features such as shadowing, power distortion, and multipath effect.

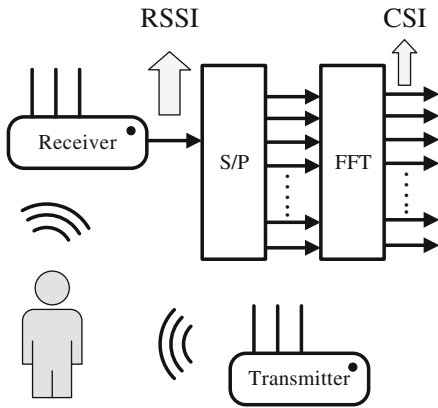
Let H_i be the CSI value of subcarrier i , which is a complex value, as

$$H_i = |H_i| \exp(j \angle H_i). \quad (1)$$

where $|H_i|$ and $\angle H_i$ are the amplitude response and phase response of subcarrier i , respectively.



Sleep Monitoring Using WiFi Signal, Fig. 1 Sleep monitoring using WiFi signals



Sleep Monitoring Using WiFi Signal, Fig. 2 CSI and RSSI in OFDM based WiFi systems

The measured phase of subcarrier i , $\angle \hat{H}_i$, is Wang et al. (2017b,c)

$$\angle \hat{H}_i = \angle H_i + (\lambda_p + \lambda_s)m_i + \lambda_c + \beta + Z, \quad (2)$$

where m_i is the subcarrier index of subcarrier i , β is the initial phase offset at the phase-locked loop (PLL), Z is the measurement environment noise, and λ_p , λ_s , and λ_c are the phase errors from the packet boundary detection (PBD), the sampling frequency offset (SFO), and the central frequency offset (CFO), respectively.

The measured CSI phase data is not stable because of the randomness in the RF chain, but the measured *phase difference* (i.e., the difference of phases from two adjacent antennas) can be exploited for vital sign monitoring, which is given by

$$\Delta \angle \hat{H}_i = \Delta \angle H_i + \Delta \beta + \Delta Z, \quad (3)$$

where $\Delta \angle H_i$ is the true phase difference of subcarrier i , $\Delta \beta$ is the unknown difference in phase offsets, which is a constant, and ΔZ is the noise difference. Obviously, $\Delta \angle \hat{H}_i$ is a stable signal, because the randomness from PBD, SFO, and CFO are all canceled.

On the other hand, RSSI can be described as a function of sender-receiver distance d with a path loss model, as in Wang et al. (2014)

$$\text{RSSI} = P_0 - 10n \log_{10}(d) + w, \quad (4)$$

where P_0 is the RSS value at a reference distance d_0 ; n is the path loss exponent (which is different from 2 in free space to 4 (or a higher value) in indoor environments), and w is a zero mean Gaussian noise. Measurements show that RSSI not only fluctuates over distance but also varies over time. Thus, a small change of the propagation environment will cause large changes in RSSI, which can also be exploited for vital sign monitoring.

Key Applications

RSS-Based Vital Sign Monitoring

UbiBreathe is a WiFi RSS-based technique, which can capture the human respiration signal with RSS values obtained from off-the-shelf WiFi devices (Abdelnasser et al. 2015). In particular, the system uses a cellphone as receiver and a laptop as transmitter. To extract the respiration signal, raw RSS values are firstly filtered by a band-pass filter with cutoff frequency from 0.1 to 0.5 Hz. Then, noise is further removed by a discrete wavelet transform (DWT) and outlier removal. Breathing rate can be then estimated using FFT. The system also detects abnormal breathing symptoms such as apnea using a threshold method, while the subject is sleeping. Experimental results show that the estimation error of breathing rate of UbiBreathe is less than 1 beats per minute (bpm). However, the system has special location requirements for the user and two devices. The system can only detect human breathing signal when the receiver is on the user's chest or the user is in the LOS path of the transmitter and receiver.

However, RSS values collected from 2.4 to 5 GHz WiFi devices can hardly capture the human heart rate because of the low resolution of RSS values. In addition, the received RSS values can be easily corrupted by other persons nearby, because of the omnidirectional propagation of WiFi signals. To address this issue, the mmVital system employs 60 GHz WiFi devices to measure both breathing rate and heart rate (Yang

et al. 2016). Since the 60 GHz WiFi signal is transmitted in a highly directional beam, the subject can be easily separated from other unrelated objects/people in the environment. In fact, it is necessary for the system to first identify the location of the user, before monitoring vital signs. To deal with this challenge, the authors develop a novel human finding technique to differentiate the reflections from the user and other surrounding objects. With this technique, the system can locate two users and monitor their vital signs simultaneously. The evaluation of the system shows that the accuracy of human finding subsystem is up to 98.4% and the mean errors of breathing rate and heart rate estimations are 0.43 and 2.15 bpm, respectively. The accuracy of the system is high, but there are still some deficiencies for the system. For example, the user should stay in the LOS or the reflection path of the transmitter and receiver, and the special 60 GHz hardware is required for this technique.

CSI-Based Vital Sign Monitoring

CSI can provide much more information on the wireless channel than RSS. Several CSI-based systems have been proposed for more accurate vital sign monitoring. The amplitude of CSI is firstly used to detect breathing rate. The WiFi-sleep system monitors human respiration and different sleeping postures by analyzing the amplitude of CSI (Liu et al. 2014). To denoise the CSI amplitude and remove outliers, the raw received signals from all subcarriers are firstly filtered by DWT. Because the sensitivities of CSI values over subcarriers are different, some of the signals may not be useful. The system only chooses the most sensitive signal to estimate the human breathing rate. Furthermore, the system also analyzes the effect of human sleeping postures on breathing estimation. To reduce the impact of different sleeping postures, the system will automatically choose the best antenna pair to obtain the most appropriate CSI signal.

In addition to breath monitoring, another work can also estimate heart rate from CSI amplitude (Liu et al. 2015). Different from the breath-

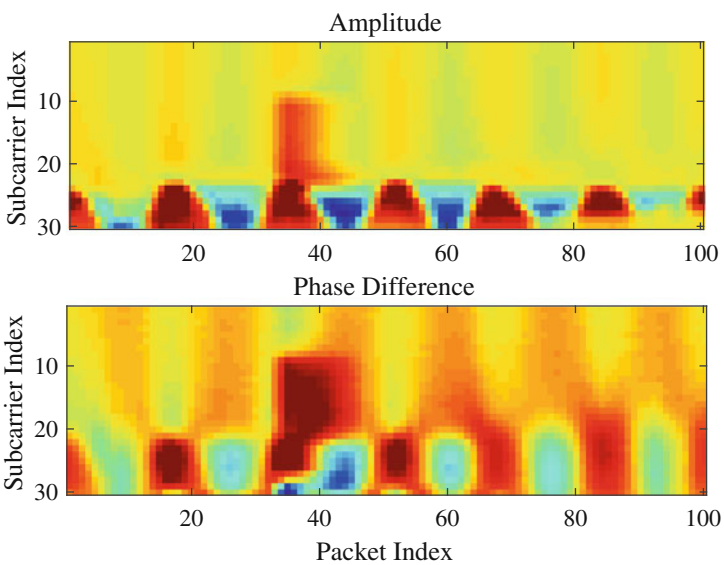
ing signal, which can be obviously observed in CSI, the heart signal is very weak and cannot be easily detected from raw CSI measurements. To focus on heart rate monitoring, the received signal is filtered by a band-pass filter with the cut-off frequency from 1 to 1.33 Hz. The experiment results with this technique show that the mean estimation errors of heart rate and breathing rate is 2 and 0.3 bpm, respectively.

Compared with CSI amplitude, CSI phase is more robust for monitoring vital signs at different distances and orientations. However, due to the asynchronous options of the transmitter and receiver, the measured phase information cannot be directly used. To this end, the PhaseBeat system leverages the phase difference between two receiving antennas to monitor vital signs (Wang et al. 2017b). With phase difference data, the robustness of respiration rate estimation is greatly increased. In addition, PhaseBeat can detect the breathing signal at longer distances, e.g., when the transmitter is 11 m away from the receiver, the median error of breathing rate estimation is 0.25 bpm. PhaseBeat can also detect heart rate with a 1 bpm median error.

Although distance has less impact on phase difference, the sensitivities of CSI phase difference and amplitude can be affected by the reflection from surroundings or the posture of the user (Wang et al. 2016b). Figure 3 shows a colormap for breathing signals using CSI phase difference and amplitude, which illustrates the independence of the sensitivities of CSI phase difference and amplitude. To enhance the robustness of CSI-based systems, the bimodal data system ResBeat leverages not only phase difference but also amplitude (Wang et al. 2017e). ResBeat implements an adaptive signal selection algorithm to ensure that only the most sensitive amplitude and phase difference be used for extracting respiration signal. The final estimation result is calculated with the combination of both amplitude and phase difference data. The system is shown to remain robust, even though the location or the posture of the user is changed. The success rate of human respiration rate estimation is higher than 90% in the experiments.

Sleep Monitoring Using WiFi Signal, Fig. 3

Colormap for breathing signals using CSI phase difference and amplitude



Sleep Monitoring Using WiFi Signal, Table 1 Comparison of different systems for vital sign monitoring

	Breathing rate	Heart rate	Multiple users	Special hardware	LOS path requirement
UbiBreathe (Abdelnasser et al. 2015)	✓	×	✓	×	✓
mmVital (Yang et al. 2016)	✓	✓	✓	✓	✓
Wi-sleep (Liu et al. 2014)	✓	×	×	×	×
Amplitude-based method (Liu et al. 2015)	✓	✓	✓	×	×
PhaseBeat (Wang et al. 2017b)	✓	✓	✓	×	×
ResBeat (Wang et al. 2017e)	✓	×	×	×	×
TensorBeat (Wang et al. 2017c)	✓	×	✓	×	×
TR-BREATH (Chen et al. 2018)	✓	×	✓	×	×

The TensorBeat system is proposed to handle the case of multiple persons with tensor decomposition (Wang et al. 2017c). First, TensorBeat uses CSI phase difference data between pairs of antennas to create CSI tensors, and then applies canonical polyadic (CP) decomposition to decompose the tensor. In addition, a stable signal matching algorithm is utilized for obtaining the decomposed signal pairs for each person. Finally, a peak detection method is exploited to estimate the breathing rates for multiple persons. Experimental results show that TensorBeat can achieve high accuracy under different environments for multi-person breathing rate monitoring. In addition, TR-BREATH is also proposed for monitoring breathing rates for multiple persons (Chen et al.

2018). CSI values is used for obtaining time reversal resonating strength (TRRS) features, which can be thus analyzed by root-MUSIC and affinity propagation algorithms to estimate multiple persons breathing rates.

Table 1 provides a comparison of different systems for vital sign monitoring in detail.

Conclusions

We addressed the problem of sleep monitoring using WiFi signals in this article. The preliminaries of WiFi signals is first introduced. Then, an overview of existing techniques of vital sign monitoring using WiFi signals is provided, where RSS- and CSI-based techniques are discussed.

Acknowledgments This work was supported in part by NSF under Grant CNS-1702957.

References

- Abdelnasser H, Harras KA, Youssef M (2015) Ubibreathe: a ubiquitous non-invasive wifi-based breathing estimator. In: Proceedings of IEEE MobiHoc'15, Hangzhou, pp 277–286
- Adib F, Mao H, Kabelac Z, Katabi D, Miller R (2015) Smart homes that monitor breathing and heart rate. In: Proceedings of ACM CHI'15, Seoul, pp 837–846
- Aly H, Youssef M (2016) Zephyr: ubiquitous accurate multi-sensor fusion-based respiratory rate estimation using smartphones. In: IEEE INFOCOM 2016-The 35th annual IEEE international conference on computer communications. IEEE, pp 1–9
- Chen Z, Lin M, Chen F, Lane ND, Cardone G, Wang R, Li T, Chen Y, Choudhury T, Campbell AT (2013) Unobtrusive sleep monitoring using smartphones. In: 2013 7th international conference on pervasive computing technologies for healthcare (PervasiveHealth). IEEE, pp 145–152
- Chen C, Han Y, Chen Y, Lai H-Q, Zhang F, Wang B, Liu KJR (2018) TR-BREATH: time-reversal breathing rate estimation and detection. IEEE Trans Biomed Eng 65 (3):489–501
- Droitcour A, Boric-Lubecke O, Kovacs G (2009) Signal-to-noise ratio in Doppler radar system for heart and respiratory rate measurements. IEEE Trans Microw Theory Technol 57(10):2498–2507
- Halperin D, Hu WJ, Sheth A, Wetherall D (2010) Predictable 802.11 packet delivery from wireless channel measurements. In: Proceedings of ACM SIGCOMM'10, New Delhi, pp 159–170
- Hou Y, Wang Y, Zheng Y (2017) Tagbreathe: monitor breathing with commodity rfid systems. In: 2017 IEEE 37th international conference on distributed computing systems (ICDCS). IEEE, pp 404–413
- Hunt C, Hauck F (2006) Sudden infant death syndrome. Can Med Assoc J 174(13):1309–1310
- Liu X, Cao J, Tang S, Wen J (2014) Wi-sleep: contactless sleep monitoring via WiFi signals. In: 2014 IEEE real-time systems symposium (RTSS). IEEE, pp 346–355
- Liu J, Wang Y, Chen Y, Yang J, Chen X, Cheng J (2015) Tracking vital signs during sleep leveraging off-the-shelf WiFi. In: Proceedings of ACM Mobihoc'15, Hangzhou, pp 267–276
- Mogue MLR, Rantala B (1988) Capnometers. J Clin Monit 4:115–121
- Mohamed R, Youssef M (2017) Heartsense: ubiquitous accurate multi-modal fusion-based heart rate estimation using smartphones. Proc ACM Interactive Mobile Wearable Ubiquit Technol 1(3):97
- Nandakumar R, Gollakota S, Watson N (2015) Contactless sleep apnea detection on smartphones. In: Proceedings of the 13th annual international conference on mobile systems, applications, and services. ACM, pp 45–57
- Nguyen P, Zhang X, Halbower A, Vu T (2016) Continuous and fine-grained breathing volume monitoring from AFAR using wireless signals. In: Proceedings of IEEE INFOCOM'16, San Francisco
- Pantelopoulous A, Bourbakis NG (2010) A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans Syst Man Cybern Part C (Appl Rev) 40(1):1–12
- Salmi J, Molisch AF (2011) Propagation parameter estimation, modeling and measurements for ultrawideband MIMO radar. IEEE Trans Microw Theory Technol 59 (11):4257–4267
- Shariati NH, Zahedi E (2005) Comparison of selected parametric models for analysis of the photoplethysmographic signal. In: Proceedings of 1st IEEE conference on computer, communication, signal processing, Kuala Lumpur, pp 169–172
- Sun X, Qiu L, Wu Y, Tang Y, Cao G (2017) Sleepmonitor: monitoring respiratory rate and body position during sleep using smartwatch. Proc ACM Interactive Mobile Wearable Ubiquit Technol 1(3):104
- Wang X, Mao S, Pandey S, Agrawal P (2014) CA2T: cooperative antenna arrays technique for pinpoint indoor localization. Proc Comput Sci 34:392–399
- Wang X, Gao L, Mao S (2016a) CSI phase fingerprinting for indoor localization with a deep learning approach. IEEE Internet Things J 3(6):1113–1123
- Wang H, Zhang D, Ma J, Wang Y, Wang Y, Wu D, Gu T, Xie B (2016b) Human respiration detection with commodity wifi devices: do user location and body orientation matter? In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing, pp 25–36. ACM
- Wang X, Huang R, Mao S (2017a) SonarBeat: sonar phase for breathing beat monitoring with smartphones. In: Proceedings of ICCCN 2017, Vancouver. IEEE, pp 1–8
- Wang X, Yang C, Mao S (2017b) PhaseBeat: exploiting CSI phase data for vital sign monitoring with commodity WiFi devices. In: Proceedings of IEEE ICDCS 2017, Atlanta. IEEE, pp 1230–1239
- Wang X, Yang C, Mao S (2017c) Tensorbeat: tensor decomposition for monitoring multiperson breathing beats with commodity WiFi. ACM Trans Intell Syst Technol (TIST) 9(1):8:1–8:27
- Wang X, Gao L, Mao S, Pandey S (2017d) CSI-based fingerprinting for indoor localization: a deep learning approach. IEEE Trans Veh Technol 66(1):763–776
- Wang X, Yang C, Mao S (2017e) ResBeat: resilient breathing beats monitoring with realtime bimodal CSI data. In: Proceedings of IEEE GLOBECOM 2017, Singapore, pp 1–6
- Wang X, Wang X, Mao S (2018a) RF sensing for internet of things: a general deep learning framework. IEEE Commun Mag 56(8):62–67
- Wang B, Xu Q, Chen C, Zhang F, Liu KJR (2018b) The promise of radio analytics: a future paradigm of wireless positioning, tracking, and sensing. IEEE Sig Process Mag 35(3):59–80
- Xie Y, Li Z, Li M (2015) Precise power delay profiling with commodity WiFi. In: Proceedings of ACM Mobicom'15, Paris, pp 53–64

- Yang Z, Pathak P, Zeng Y, Liran X, Mohapatra P (2016) Monitoring vital signs using millimeter wave. In: Proceedings of IEEE MobiHoc'16, Paderborn, Germany, pp 211–220
- Yang C, Wang X, Mao S (2018) AutoTag: Recurrent variational autoencoder for unsupervised apnea detection with RFID tags. In: Proceedings of IEEE GLOBE-COM'18, Abu Dhabi, United Arab Emirates, pp 1–7

Smart Agriculture

- [Ubiquitous Big Data](#)

Smart City

- [Architecture and Data Management for Smart Community Information Platform](#)

Smart Community

- [Architecture and Data Management for Smart Community Information Platform](#)

Smart Community Services

- [Architecture and Data Management for Smart Community Information Platform](#)

Smart Device-to-Device Communication: Communication Establishment and Maintenance

Shaoyi Xu
Department of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

Synonyms

[Device-to-Device communication](#); [Hybrid system](#); [Long-Term Evolution-Advanced](#); [Near-far interference](#)

Definitions

Unlike the infrastructure-based cellular network, D2D users (user equipment or mobile terminals) do not communicate via the central coordinator (base station, NodeB, or evolved NodeB) but operate as an underlay and communicate directly with each other or more hops.

Historical Background

A new generation of wireless communications, the fifth generation (5G), is predicted for beyond 2020. For 5G, much higher data rate and better user experience are two most obvious requirements (Saxena et al. 2020). To satisfy the first requirement, device-to-device (D2D) communication has been proposed, which allows two users in proximity communicate directly rather than go through the base station (BS). Correspondingly, three promising gains, i.e., proximity gain, reuse gain, and hop gain, can be offered to improve system performance (Han et al. 2019). Meanwhile, the gains can be achieved without having to invest in network-side hardware upgrades or new cell deployments.

Actually, the trait of D2D communication is not new in the cellular system. In academia, D2D communication was first proposed in (Lin and Hsu 2000) to enable multihop relays in cellular networks. Later the works investigated the potential of D2D communications for improving spectral efficiency of cellular networks, multicasting, peer-to-peer communication, video dissemination, machine-to-machine (M2M) communication, cellular offloading, content catching, and so on (Chatzopoulos et al. 2019). D2D also has been identified in some specifications such as Private Radio Communication System (PRCS), Japanese Handy Phone System (HPS), Digital Enhanced Cordless Telecommunications (DECT), etc. Many companies devote actively much in this research. Qualcomm, for example, is heading and has been very active recently. A prototype of “FlashLinQ” modem has been developed since 2006 to enable direct D2D communication

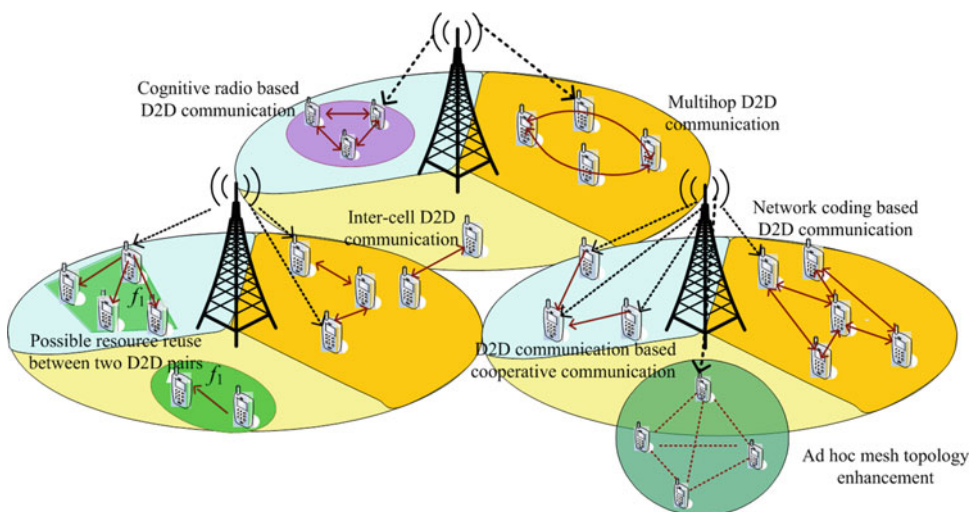
over a licensed spectrum. The prototype effort covers the protocol stack from the host software interface to the physical layer algorithm. “FlashLinQ” is even a new trademark registered in 2008 by Qualcomm (QUALCOMM 2008).

The key motivation for utilizing D2D communication underlying a cellular network is to keep local communication local. D2D communication shares the same resources with the cellular system while under the control of the evolved NodeB (eNB). In such a case, an eNB can still control the resource and power assigned for D2D transmission to limit the interference to the primary cellular networks. Although the concept of D2D communication as an underlay to a cellular network is not new in a cellular system, it is new in Long-Term Evolution (LTE) systems. As a result, some new challenges are incurred. The first key issue to be resolved is how to facilitate D2D communication in the licensed frequency band. Different from a centralized system where an eNB fully controls the data transmission which should be sent on the pre-allocated resources and user equipment (UE) terminals must synchronize with the eNB, devices would compete with each other for the resource in a distributed way in the D2D subsystem. Secondly, interference management in a hybrid network is a critical issue in that

the interference from D2D links can be expected to reduce the cellular capacity and efficiency and vice versa (Xu et al. 2012).

Foundations

- **System Model:** Indeed, such D2D communication is likely to become integral to the future beyond 3G world to form a hybrid network. Fig. 1 illustrates the basic and extended concept of D2D communication and hybrid networks. In Fig. 1, a UE device can communicate with another UE device directly or by the aid of a base station. Moreover, multiple D2D subsystems can share resources in a cell.
- **Used Spectrums:** Generally speaking, the spectrum that D2D communication may use can be categorized into inband and outband.
 - **Inband D2D:** The spectrums are used for both D2D system and cellular network. For the case that these two systems share same spectrum, it is called underlay inband D2D, and thus interference is incurred. In order to avoid the aforementioned interference issue, some researchers propose to dedicate part of the cellular resources only to D2D communications (i.e., overlay



Smart Device-to-Device Communication: Communication Establishment and Maintenance, Fig. 1 Illustration of D2D communication as an underlay to a cellular system

inband D2D). However, dedicated cellular resources may be wasted.

- Outband D2D: To eliminate the interference issue between D2D and cellular link, the D2D links exploit unlicensed spectrum. However, using unlicensed spectrum requires an extra interface and usually adopts other wireless technologies such as Wi-Fi Direct, ZigBee, or Bluetooth. In outband communications, the coordination between radio interfaces is controlled by either the BS (i.e., controlled) or the users themselves (i.e., autonomous). Outband D2D communication faces a few challenges in coordinating the communication over two different bands because usually D2D communication happens on a second radio interface.

Nevertheless, for the case of outband D2D, the interference in unlicensed spectrum is difficult to control. Especially, only cellular devices with two radio interfaces (e.g., LTE and Wi-Fi) can use outband D2D communications. More important, packets (at least the headers) need to be decoded and encoded because the protocols employed by different radio interfaces are not the same. Based on the above reasons, inband D2D has obtained wider research.

- Communication Establishment: Proximate device discovery, or proximate peer discovery, as defined by Qualcomm means the ability for a device to passively and continuously search for relevant value in one's physical proximity. Actually, it can be further identified in terms of radio interface and networking applications.
 - Discovery of radio interface feature: A device is able to discover and be discovered by other devices in radio proximity.
 - Discovery of application layer feature: A user of a service or social networking application is able to discover and be discovered by other users of the same service or social networking application.

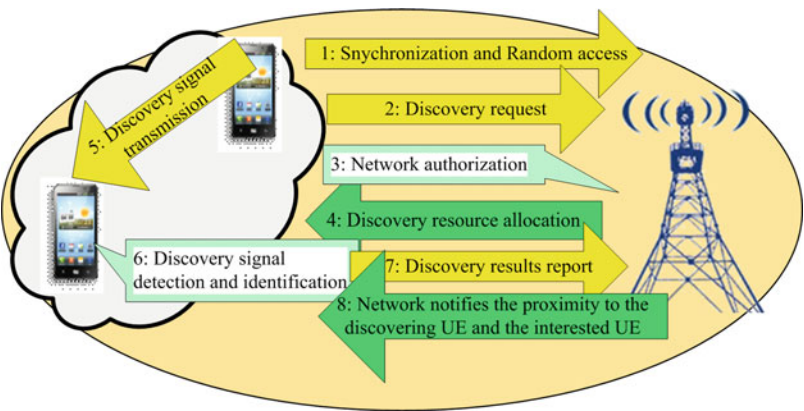
In generally, D2D discovery involves three procedures, namely, signature discovery, identity discovery and verification, and service discovery. Signature discovery means

the device detection includes acquiring the transmission frame timing and synchronization of the D2D receiver with the D2D transmitter. Identity discovery and verification means the procedure that the source device can retrieve the identity of target devices through the discovered signature obtained from the discovery channel. Service discovery identifies the desired services of D2D communications after the target devices are detected and verified. The general requirements for device discovery are no doubt that all of the terminals want to be discovered fast by more neighbors with low energy consumption, whereas the legacy impact, including the additional interference to the current wide area network (WAN) and usable resource degradation caused by D2D discovery, should be minimized. Both infrastructure and ad hoc networks can be deployed for peer discovery. With the lack of network support, peer discovery in a distributed situation is time- and energy-consuming since a great deal of beacon signals and sophisticated scanning are required. Moreover, interaction signaling from higher layers between two end users will be involved to implement security procedures. Contrarily, many benefits may be obtained from the infrastructure such as synchronization, resource allocation, and interference coordination and thus make the discovery procedure more efficient. Consequently, studies on the network controlled discovery are explicitly solicited by the 3GPP study item for Rel-12 (3GPP 2013). Figures 2 and 3, respectively, illustrate the network-directed device discovery procedures and peer discovery procedures in LTE Direct.

- Interference Management: For the case of inband D2D, interference management in such a hybrid network is a critical issue in that the interference from D2D links is expected to reduce the cellular capacity and efficiency. Interference may occur when D2D devices share uplink (UL) or downlink (DL) resources with cellular systems.

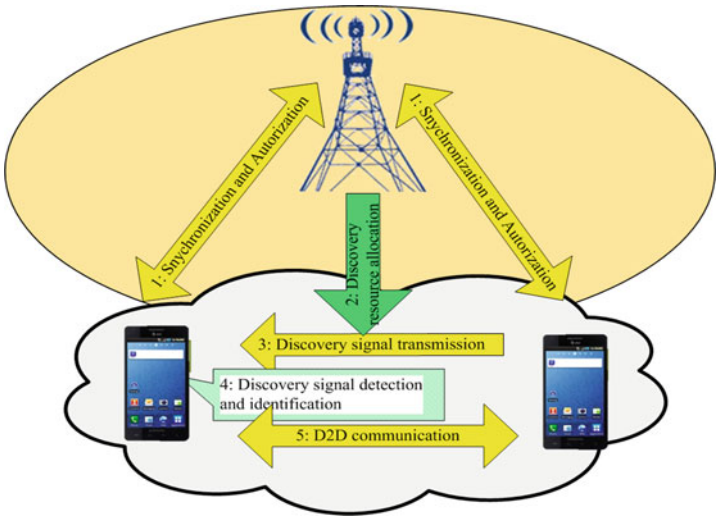
Smart Device-to-Device Communication: Communication Establishment and Maintenance, Fig. 2

Network-directed device discovery procedures



Smart Device-to-Device Communication Establishment and Maintenance, Fig. 3

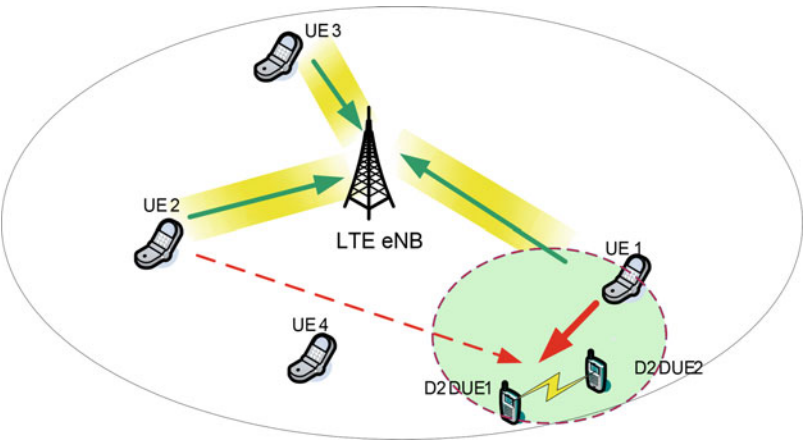
Peer discovery procedures in LTE Direct



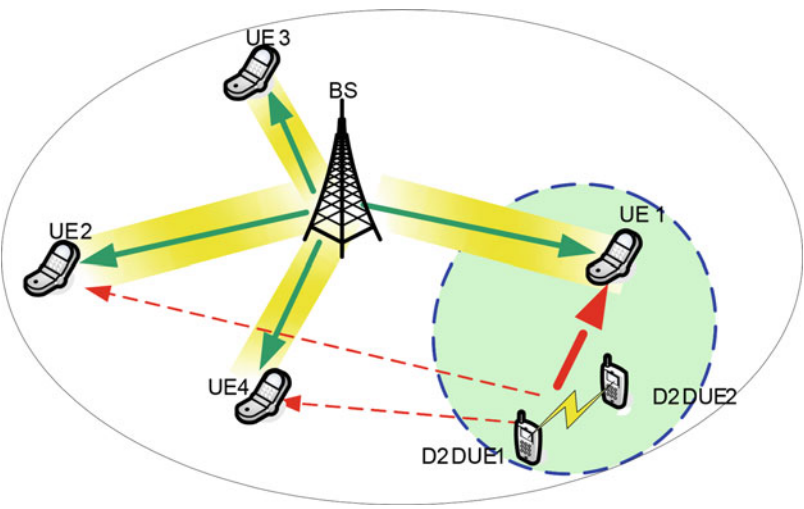
- In the LTE UL, a cellular UE device sends data after several TTIs on receipt of resource allocation information from the eNB. Therefore, D2D UE may finish its RRM by using the related cellular information within these several milliseconds. Additionally, the UL is underutilized by the cellular operators comparing to DL. Thereby, most research of D2D transmission focuses on the scenario that D2D users use the uplink spectrum of an LTE system. Fig. 4 shows the near-far interference where cellular UE1 may impose serious interference to D2D transmission if they are sharing the same resource, whereas the interference from UE2 can be negligible since the long distance between the D2D pair and UE2. To arrange the dedicated

- resource for the cellular system and D2D users is an intuitive solution but leads to inefficient utilization of the available resources and low system throughput.
- In an LTE-like cellular network, since continuous control signaling and data are transmitted and fast scheduling are used in the DL spectrum, it is very difficult for the D2D transmission to share DL spectrums with cellular users. Other than the fast resource allocation, possible near-far interference is another serious problem when DL spectrums are reused, and in Fig. 5, we observe that the D2D UE1 and D2D UE2 may impose significant interference to the cellular UE1 since the short distance between them. So the key problem for sharing DL spectrums of a

Smart Device-to-Device Communication: Communication Establishment and Maintenance, Fig. 4 An example of near-far interference in the LTE uplink spectrum



Smart Device-to-Device Communication: Communication Establishment and Maintenance, Fig. 5 Near-far interference when sharing cellular downlink spectrums



- hybrid system is to avoid interference to nearby cellular UE devices.
- Many interference management methods can be utilized to coordinate the near-far interference, such as transmission mechanism design (Xu et al. 2010, 2012), power control (Xu et al. 2015, 2016), resource allocation (Xu et al. 2016), transmission mode selection (Chen et al. 2018), and so on.

related industries such as content caching, mobile commerce, mobile advertising, mobile security, and cloud services to prosperity. It is also basic and expended applications of vehicle-to-vehicle (V2V) communications and machine-to-machine (M2M) communications.

Key Applications

D2D communications are important for location-based services in current 5G networks, leading

Cross-References

- ▶ [Device-to-Device \(D2D\) Communication](#)
- ▶ [Interference Management](#)
- ▶ [Machine to Machine \(M2M\)](#)
- ▶ [Vehicle-to-Vehicle Communications](#)

References

- 3GPP TR 22.803, Feasibility study for Proximity Services (ProSe), V12.2.0, June 2013
- Chatzopoulos D, Bermejo C, Haq E-U et al (2019) D2D task offloading: a dataset-based Q&a. *IEEE Commun Mag* 57(2):102–107
- Chen C, Sung C, Chen H (2018) Optimal mode selection algorithms in multiple pair device-to-device communications. *IEEE Wirel Commun* 25(4): 82–87
- Han B, Sciancalepore V, Holland O et al (2019) D2D-based grouped random access to mitigate mobile access congestion in 5G sensor networks. *IEEE Commun Mag* 57(9):93–99
- Lin Y-D, Hsu Y-C (2000) Multihop cellular: a new architecture for wireless communications. In: *Proceedings of 19th international conference on computer communications Israel*, March 2000, vol 3, pp 1273–1282
- QUALCOMM Incorporated (2008) FLASHLinq Trademark Information. <http://www.trademarkia.com/flashling-77422968.html>. Accessed 14 Mar 2008
- Saxena N, Kumbhar F-H, Roy A (2020) Exploiting social relationships for trustworthy D2D relay in 5G cellular networks. *IEEE Commun Mag* 58(2): 48–53
- Xu S, Wang H, Chen T et al (2010) Effective interference cancellation scheme for device-to-device communication underlying cellular networks. In: *Proceedings of 72th IEEE VTC-Fall Ottawa*, Sept 2000, pp 1–5
- Xu S, Wang H, Chen T et al (2012) Device-to-device communication Underlying cellular networks: connection establishment and interference avoidance. *KSII Trans Internet Inf Syst* 6(1):203–228
- Xu S, Kwak KS, Rao R (2015) Interference-aware resource sharing in D2D underlying LTE-A networks. *Trans Emerg Telecommun Technol* 26(12):1306–1322
- Xu S, Xia C, Kwak KS (2016) Overlapping coalition formation games based interference coordination for D2D underlying LTE-A networks. *AEU-Int J Electron Commun* 70(2):204–209

Smart Grid Communication Based on IEEE 2030 Standard

Zied Bouida¹, Javad Fattahi², Arslan Ahmed¹, Mohamed Ibnkahla¹, Henry Schriemer², and Raed Abdullah³

¹Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

²University of Ottawa, Ottawa, ON, Canada

³Hydro Ottawa, Ottawa, ON, Canada

Synonyms

Reliable; Scalable smart grid communication; Secure; Smart energy profile 2.0.; Two-way

Definitions

The Grid refers to the electric grid that can be defined as a network of several components including transmission lines, stations, and transformers aiming at delivering electricity from the power generators to end users. A Smart Grid is a system that uses two-way communication to enable bidirectional energy flows, in particular, within active distribution networks. It also enables sensing the transmission lines to achieve a more efficient electricity management.

The IEEE 2030 Standard defines a set of specifications that exploit the Smart Grid interoperability reference model and framework in power applications, DC quick charger, energy storage system, information exchange, and control through communications.

Historical Background

Energy utilities are constantly under pressure to meet the growing and complicated energy demands (Amin and Wollenberg 2005; Farhangi 2010). The traditional energy grid allows for

Smart Energy Profile 2.0

- [Smart Grid Communication Based on IEEE 2030 Standard](#)

one-way communication of energy usage between users and the utility. This does not allow utilities to have control or to suggest any changes in consumption based on the energy consumption data they obtain. In this context, the risks associated with unidirectional systems have grown and their sustainability has become a concern. These risks come with higher costs and degraded quality of service including unscheduled power outages and blackouts. To address this, smart grid systems are implemented to allow a two-way communication between the utility and its customers (Fang et al. 2012).

With advances in smart grid communications, many of the challenges discussed above have been resolved. However, given the heterogeneity of devices and protocols used in smart grid communications and the lack of interoperability between different grid's components, the need for a standard addressing these issues became apparent. To this end, IEEE has published IEEE 2030 as a standard for smart grids (IEEE 2013). Under this standard, IEEE 2030.5, also called Smart Energy Profile 2.0, defines the communication between consumers and the Utility. The standard also grants consumers the ability to manage their energy consumption. Moreover, it helps in the integration of smart devices and smart applications by defining a framework of secure and interoperable communication.

In light of the above, the details behind the implementation of a smart grid communication system compliant with the IEEE 2030 standard are given in this chapter. In what follows, an overview of the IEEE 2030 and its features is first given. Then, the implementation of the communication system in compliance with the standard is provided. After that, data analytics are illustrated as additional tools to further enhance the Smart Grid system performance.

Foundations

- Overview of the IEEE 2030 Standard and its Features

The IEEE 2030.5 standard was proposed to address the lack of interoperability in cur-

rent smart grid systems. Indeed, smart grid interoperability is a necessary foundation in order to provide seamless and end-to-end system connectivity. Smart grid interoperability is dealing with (1) power systems, (2) communications, and (3) information technology. The details behind these interoperability architectural perspectives (IAP) are given below.

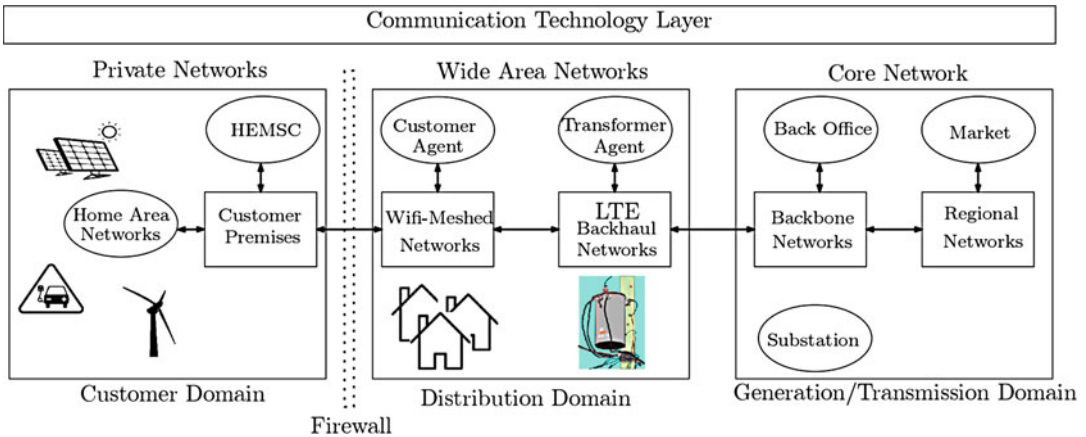
- Power systems IAP (PS-IAP): The emphasis of the smart grid interoperability reference model (SGIRM) in terms of power includes apparatus, applications, and operational concepts. This perspective defines seven domains – bulk generation, transmission, distribution, service providers, markets, control/operations, and customers – to enable both traditional and alternative smart grid technology approaches in energy balance, equipment rating and optimization, and localized (distributed) control.
- Communications technology IAP (CT-IAP): The focus of SGIRM CT-IAP is communication connectivity among systems, devices, and applications in the context of the Smart Grid. The communication media between all seven domains may include wide area networks (WAN), neighborhood area networks (NAN), field area networks (FAN), and home area networks. The communications entities, wireline or wireless, stand in the system architecture, but the interfaces between two or more entities are defined as generic interconnections that establish the minimum level of interoperability. Thus, interfaces are specified in terms of performance requirement, security level, protocol layer, and other more specific needs.
- Information technology IAP (IT-IAP): SGIRM IT-IAP provides the context for information exchange interoperability among smart grid applications by the control of processes and data management flow. Designing data models represents a crucial part in structuring an information technology architecture and indicates the

intact governance process. Establishing common semantics in data model facilitates data use by multiple applications. IEEE 2030 uses a Common Information Model (CIM) from IEC 61970 and IEC 61968 to enable interoperability of the data exchanges. This happens through the use of standardized object models that provide the abstract information model (semantic), profiles for standardization (contextual), and the schemas views (syntactical). Moreover, it brings inter-application data exchange to the distributed application programming interface (API), and – within a utility’s enterprise system – of the information elements used within the smart grid infrastructure. It uses an ontology-based strategy to create and manipulate the data model, which provides easy translation and promotes interoperability by using Extensible Markup Language (XML) or Unified Modeling Language (UML). Using ontology in data model provides a formal semantics based on a shared understanding that is machine-readable.

In order to address the above interoperability features, the IEEE 2030 standard encompasses the following standards:

- IEEE Std 2030.1: Covers the technical specifications of a DC quick charger for the use of electric vehicles (EV) and original equipment maker (OEM).
- IEEE Std 2030.2: Applies SGIRM process to energy storage by emphasizing on the information relevant to ESS interoperability with EPS.
- IEEE Std 2030.3: Establishes the tests and procedures to verify if ESS meet the safety and reliability requirements of the EPS.
- IEEE Std 2030.5: This standard represents the main focus of this chapter and provides a reference document for Smart Energy Profile Protocol version 2.0 (SEP 2.0). It defines an application profile interfacing the smart grid and users enabling them to manage distributed generation, electric vehicles, energy storage, demand response, load control, and price communication. It can also be applicable in other commodities including water, natural gas, and steam.
- IEEE Std 2030.6: Presents a framework for evaluating the benefits of demand response (DR) programs and the calculation methods that would be useful in electricity market structures to provide utilities with the references for the planning and post-planning of DR programs.
- IEEE Std 2030.7: Provides technical specifications for the microgrid energy management system (MEMS). This standard addresses the technical challenges associated with the MEMS’s operation regardless of the topology and presents the control approaches required by the distribution system operator (DSO).
- Transactional smart grid (TSG) Communication Compliant with IEEE 2030.5

After discussing the need for a Smart Grid system and the importance of the standard, a real implementation of a transactional smart grid communication compliant with IEEE 2030.5 standard is discussed here. In this context, Fig. 1 represents the system model used for the communication between different Home Energy Management System Controllers (HEMSC) and Customer Agent (CA) as the primary grid edge resources (Customer domain), the transformer agent (TA) (Distribution domain), and the Utility headquarter (Transmission domain). As seen from this system architecture, CA is physically located at the customer premise near the service entrance but is administered by the utility. Since HEMSC is not trusted by the utility companies, the CA is making a firewall between the customer and distribution domains while verifying the DR action. The CA communicates with the TA to coordinate DR actions and exposes to the customers’ energy management system (EMS) a secure interface, so that both customer and utility privacy are preserved. TA communication with the utility central control room (CCR) provides centralized



Smart Grid Communication Based on IEEE 2030 Standard, Fig. 1 TSG communication system

grid condition monitoring and control. TA communication with HEMSCs via CA coordinates all DR requests and returns, and monitors the local microgrid on the low-voltage side of the transformer. The HEMSC is a residential unit responsible for control and scheduling of customer. It operates within a SEP 2.0 framework to address TA requests, based on customer choices of consumption mode, comfort level, and load priorities. The HEMSC translates data from such services to a SEP 2.0 recommended XML file for TDR purposes, including local power measurements so that DR action can be verified. It maintains customer privacy by design, acting as a customer agent to firewall any knowledge of load-specific activities and customer behavior from the TA. Moreover, Low-level DR events are initiated by the TA in response to local asset monitoring, such as transformer health.

The details behind this communication system and its compliance with IEEE 2030.5 are given in Ghalib et al. (2018) and summarized below:

- **Mesh Architecture for scalability:** In order to implement a scalable smart grid system, a Mesh architecture is selected (Akyildiz and Wang 2005; Akyol et al. 2010). This decentralized implementation does not rely on any

pre-existing infrastructure to coordinate the communication between the clients. It establishes peer-to-peer connections (Gharavi and Hu 2011). Each node in the network is connected to adjacent nodes and participates in routing by forwarding data. The determination of which nodes forward data is made dynamically based on their network connectivity. The decentralized communication model allows for a smoother self-healing process for the network. In this context, if one node drops for whatever reason, communication is rerouted through a different node.

- **Message Format and telemetry:** As per the IEEE 2030.5 standard, the format of the communicated files is XML and EXI. TAs and CAs are implemented as servers and clients to develop a RESTful communication used to exchange energy data and action requests. These messages have different function sets such as demand response and load control, load shed availability, consumption reduction negotiation, and price communication. Figure 2 describes the SEP 2.0 messaging sequence via GET commands for XML files. High-level DR events are initiated by the CCR server and are communicated to the TA. The TA communicates with individual CAs and consequently CA with

each individual HEMSCs, sending XML function sets with their corresponding elements in the sequence shown. HEMSCs send similar messages to the CA as part of the transactional negotiation and monitoring process, and CA sends likewise to the TA. In this architecture, TA and CA may both be either client or server, as required by the downstream devices. In this architecture, client issues a DNS-SD keyed by its short form device identification (SFDI) and request to locate its *EndDevice* payload, and then servers provides a DNS-SD response with URL to client. At this moment, client performs TLS and authentication and have an encrypted connection. Client GETs its *EndDevice* and server responds with the *EndDevice* resource that contains URL links to other payloads such as *RegistrationLink*, *FunctionSetAssignmentsListLink*, etc. After passing registration through *RegistrationLink*, and getting the finding assigned function via getting XML payloads by *FunctionSetAssignmentsListLink*, then the sequence of getting DR settings follows the steps illustrated in Fig. 2 starting from *DemandResponseProgramList* to *EndDeviceControl*. In response to *EndDeviceControl*, client must perform a response using *DrResponse*, which contains elements similar to those that are listed in the *EndDeviceControl*. In addition *LoadShedAvailability* addresses how long the consuming device is able to reduce consumption and at what maximum response level. In the case of an emergency, when prompt load shedding is needed, the aggregated data from this function set is helpful.

- Security Measures: Security is very critical to smart grid systems. Daily, a large amount of cyberattacks or breaches occur (Farraj et al. 2016). In 2015, an estimated 500 million personal information records were stolen or lost. The security measure taken into consideration in the

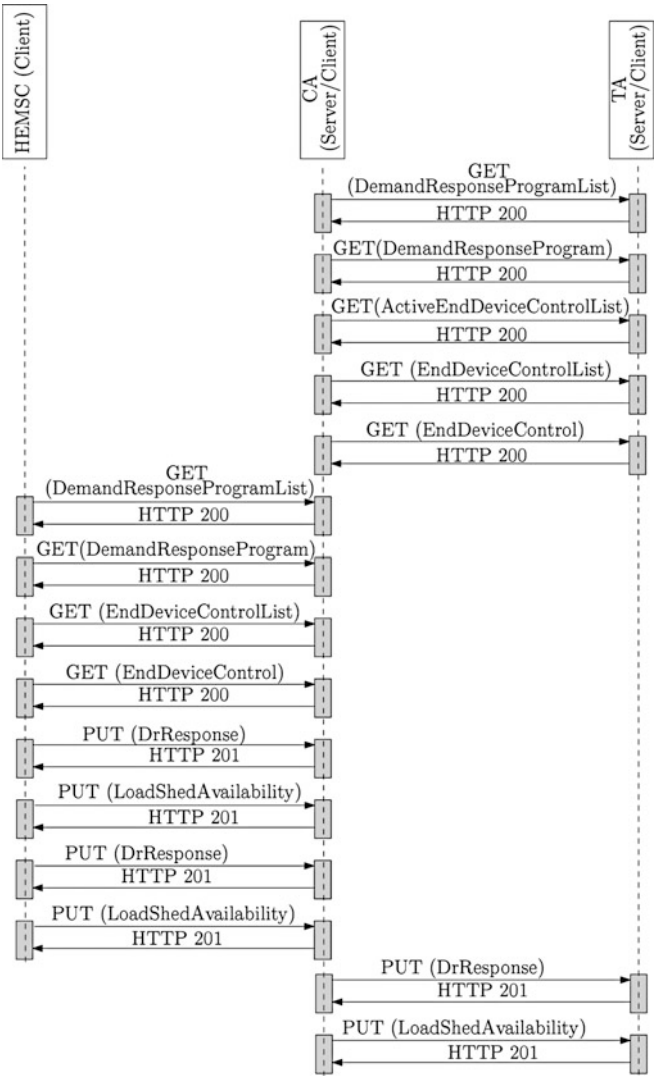
implemented smart grid communication are as follows:

Network Authentication: To protect the communication network, certain configurations can be used including unique SSIDs, static IP addresses, and passwords. The Mesh network is created after all devices boot up. The CA and TA nodes constantly look for a matched Mesh network with their static IP addresses. The TA device establishes a Mesh network under a certain SSID and excludes non-authorized IP addresses which ensures that only CAs are authorized to join the network and can communicate with other nodes in the network and prevents unauthorized entities from joining.

Encryption: A remote access feature is added to allow secure remote access using the Secure Shell protocol (SSH). While it is important to secure the end-points, it is also extremely important to secure the data flowing on the network. This is done by applying encryption using the strong asymmetric RSA with a key size of 2048. For further compliance with IEEE 2030.5 standard, TLS 1.2 is implemented.

- Security Measures: Physical protection: All TA and CAs nodes need to be well-packaged according to the IEEE 2030 standard, ensuring that they are fully protected not only from harsh weather conditions but also from theft and intruders.
- Data Analytics: After establishing secure two-way communication between different homes and the Utility as discussed above, energy consumption information is sent through this network from the CAs to the TA and then to the Utility Headquarters, as detailed in Figs. 1 and 2. Consequently, significant data may accrue at the TA which in turn relays this data to the Utility back office. Taking advantage of this

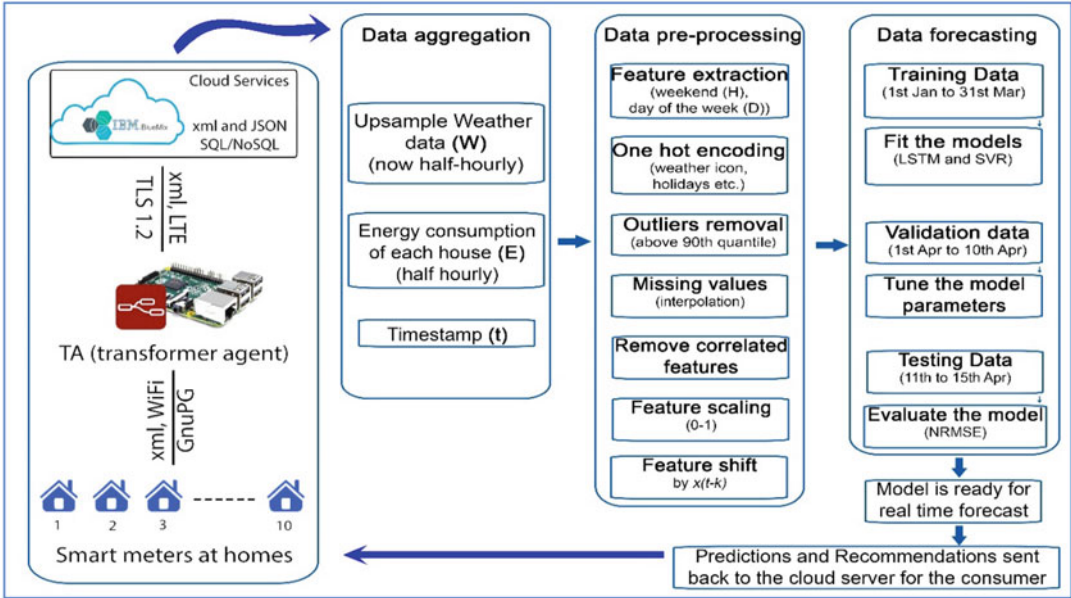
**Smart Grid
Communication Based
on IEEE 2030 Standard,
Fig. 2** Communication
hierarchy using SEP 2.0
between TA, CA and
HEMSC for TSG



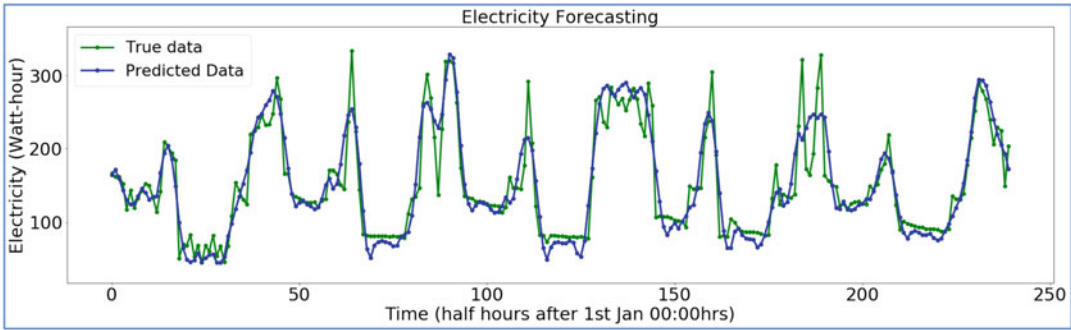
available data, analytics, and power consumption prediction are used to further enhance the performance of the implemented smart grid systems. To this end, data collected at multiple TAs is finally sent through LTE to the Cloud for storage and processing, as detailed in Fig. 3. This data undergoes aggregation and pre-processing, and is then used in forecasting. Therefore, the utility can visualize different patterns, explore and forecast energy consumption in various neighborhoods, and classify houses based on their consumption (Ahmed et al. 2018). Prediction techniques

can then be applied to forecast future consumption and identify service and asset health issues.

An example of electricity consumption forecasting is given in Fig. 4 where the energy (in Watt-hours) is predicted for the next 5 days and compared to the actual energy consumed at that time. Energy readings have been taken after every half-hour. An Autoregressive Integrated Moving Average (ARIMA) model is used for data prediction. Using the actual data, the accuracy of the prediction model is computed, and it reached 93%. The developed



Smart Grid Communication Based on IEEE 2030 Standard, Fig. 3 Data analytics processes



Smart Grid Communication Based on IEEE 2030 Standard, Fig. 4 Energy consumption forecasting using data analytics

code for data prediction is compiled on Python Jupyter Notebook.

Cross-References

- Smart Metering, Specific Challenges, and Solution Approaches
- Ultra-Reliable and Low-Latency Communications for the Smart Grid

References

Ahmed A, Arab K, Bouida Z, Ibnkahla M (2018) Data communication and analytics for smart grid systems. In: Proceedings of IEEE International Conference on Communications (ICC'2018), Kansas City, MO, pp 1–6

Akyildiz IF, Wang X (2005) A survey on wireless mesh networks. IEEE Radio Commun 43(9):23–30

Akyol B, Kirkham H, Clements S, Hadley M (2010) A survey of wireless communications for the electric power system. Prepared for the US Department of Energy

- Amin SM, Wollenberg BF (2005) Toward a smart grid: power delivery for the 21st century. *IEEE Power Energy Mag* 3(5):1540–7977
- Fang X, Misra S, Xue G, Yang D (2012) Smart grid – the new and improved power grid: a survey. *IEEE Commun Surv Tutorials* 14(4):944–980
- Farhangi H (2010) The path of the smart grid. *IEEE Power Energy Mag* 8(1):18–28
- Farraj A, Hammad E, Daoud AA, Kundur D (2016) A game-theoretic analysis of cyber switching attacks and mitigation in smart grid systems. *IEEE Trans Smart Grid* 7(4):1846–1855
- Ghalib M, Ahmed A, Al-Shiab I, Bouida Z, Ibnkahla M (2018) Implementation of a smart grid communication system compliant with IEEE 2030.5. In: *IEEE ICC 4th international workshop on integrating communications, control, and computing technologies for smart grid (ICT4SG)*. Kansas City, MO, pp 1–6
- Gharavi H, Hu B (2011) Multigate communication network for smart grid. *Proc IEEE* 99(6):1028–1045
- IEEE (2013) IEEE adoption of smart energy profile 2.0 application protocol standard. *IEEE Std 2030.5–2013*, pp 1–348

ture, which can dynamically adapt network resources for services.

Historical Background

Many popular technologies appear to be contributing to the future Internet architecture. Typically, four kinds of mainstream technologies are being widely researched currently: Information Centric Networking (ICN) (Ahlgren et al. 2012), Software Defined Network (SDN) (McKeown et al. 2008), Network Function Virtualization (NFV) (Herrera and Botero 2017), and Locator/Identifier Separation Protocol (LISP) (Fuller et al. 2013). These technologies are proposed to solve the problems of current Internet from different viewpoints.

Smart Grids

► Ubiquitous Big Data

Smart Health

► Ubiquitous Big Data

Smart Identifier Network

Hongke Zhang, Wei Quan, and Chengxiao Yu
Beijing Jiaotong University, Beijing, China

Synonyms

Future Internet; Identifier network; Innovative networking architecture

Definition

Smart Identifier Network (SINET) is a novel three layers-two domains future network archi-

- ICN: To meet users' increasing and changing demands, the communication mode should be changed from host-centric to information-centric (Xylomenos et al. 2014). ICN is based on Named Data Objects, such as web pages and videos. ICN architecture includes in-network caching and multiparty communication through replication, thus facilitating the efficient and timely delivery of information to users.
- SDN: It is proposed to decouple the control logic from data planes, enabling a logically centralized and programmable control plane. Thus, the SDN can provide more efficient configuration and higher flexibility to accommodate innovative network designs.
- NFV: It proposes to virtualize entire classes of network node functions into building blocks by evolving IT virtualization-related technologies. These virtual blocks may be flexibly connected or chained to create communication services.
- LIPS: It separates the IP address into two namespaces: Endpoint Identifiers (EIDs), which are assigned to the end hosts, and Routing Locators (RLOCs), which are assigned to the devices that make up the global routing system.

Foundations

In order to provide emerging services and improve network performance, SINET evolves to be more intelligent, dynamic, and adaptive. The network model comprises three layers, namely: *Smart Pervasive Service Layer*, *Dynamic Resource Adaption Layer*, and *Collaborative Network Component Layer*. It also contains two domains called *Entity Domain* and *Behavior Domain*. The network reference model is shown in Fig. 1.

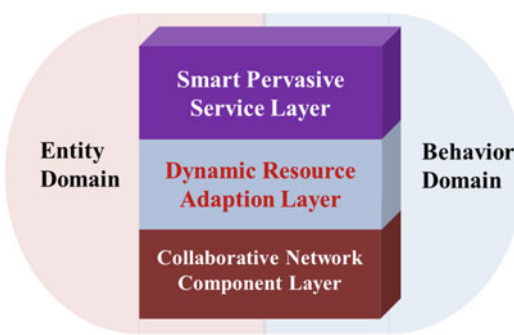
In this architecture, the Smart Pervasive Service Layer is responsible for the naming and description of services. The *Dynamic Resource Adaption Layer* dynamically adapts network resources and builds network function-groups through perceiving service demands and network status so as to satisfy the service demands and improve the quality of experience. The *Collaborative Network Component Layer* is responsible for the storage and transmission of data. Meanwhile, the *Collaborative Network Component Layer* is also responsible for the network components' behavior perceiving and aggregation. The *Entity Domain* includes the naming of service, network function-group, and component. And the *Behavior Domain* represents the description of service, network function-group, and component. In the *Dynamic Resource Adaption Layer*, the entities in the *entity domain* are generated, according to the *behavior domain* decision, which leads to the separation of control

plane from the data plane in SINET (Zhang et al. 2016).

At the same time, smart and collaborative functions have been designed into the *Smart Pervasive Service layer* and *Collaborative Network Component layer* to interact with the *Dynamic Resource Adaption layer*. With these enhanced functions, SINET can achieve the decoupling of the three bindings thoroughly. With these features, SINET further resolve the scalability, mobility (Zhang et al. 2015), security, and controllability issues which have been partially solved in SINET fundamental model. Furthermore, SINET can solve the problem of resource utilization, energy efficiency, and support of cloud computing and big data transmission.

As shown in Fig. 2, *Service Identifier (SID)* is introduced as a unique and independent namespace to name a smart service in Entity Domain, which is used to separate the resource and location of services. *Function-group Identifier (FID)* is used to name a network function-group. *Node Identifier (NID)* (Luo et al. 2014) is used to name a network component device. The Entity Domain leverages FID and NID to avoid the “control and data binding.” In the Behavior Domain, *Service Behavior Description (SBD)*, *Function-group Behavior Description (FBD)*, and *Node Behavior Description (NBD)* are used to provide further description of SID, FID, and NID, respectively. They are widely used in the enhanced mapping process in SINET which is shown in the following. Three intelligent mapping functions are proposed, namely: F1, F2, and F3. F1 chooses the appropriate network function-groups according to the service demands. F2 matches the network components of a network function-group with the service demands. F3 achieves the behavior aggregation of network components by using intra-function-group cooperation mechanism (Guan et al. 2017).

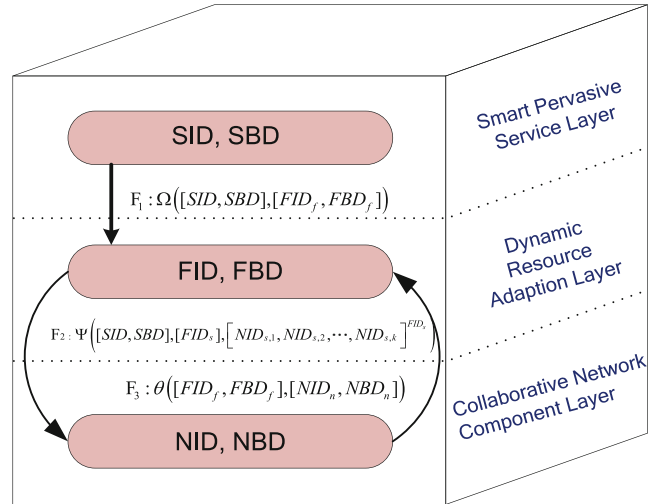
By applying the above mechanisms, emerging service types are supported with high efficiency, the utility ratio of network resource is improved, and the energy consumption of network is reduced.



Smart Identifier Network, Fig. 1 Evolutionary reference model of SINET

Smart Identifier Network, Fig. 2

Enhanced mapping models



Key Applications

Smart Identifier Network (SINET) is a specific instance of innovative networking architecture, which can promote the deployment of collaborative vehicular edge computing (Wang et al. 2018). SINET can easily accomplish optimal decision, task allocation, and resource dynamic scheduling through managing network function groups logically and can solve the problem of resource utilization, energy efficiency, and support for cloud computing and big data transmission.

Cross-References

- [Information-Centric Networking: Basic Principle and Architecture](#)
- [Network Function Virtualization](#)
- [Software-Defined Wireless Networking](#)

References

- Ahlgren B, Dannewitz C, Imbrenda C, Kutscher D, Ohlman B (2012) A survey of information-centric networking. *IEEE Commun Mag* 50(7):26–36
- Fuller F, Meyer V, Lewis D (2013) Locator/ID Separation Protocol (LISP). IETF RFC 6830. <https://tools.ietf.org/html/rfc6830>
- Guan J, Yan Z, Yao S, Xu C, Zhang H (2017) GBC-based caching function group selection algorithm for SINET. *J Netw Comput Appl* 85(2):56–56
- Herrera JG, Botero JF (2017) Resource allocation in NFV: a comprehensive survey. *IEEE Trans Netw Serv Manag* 13(3):518–532
- Luo H, Zhang H, Zukerman M, Qiao C (2014) An incrementally deployable network architecture to support both data-centric and host-centric services. *IEEE Netw* 28(4):58–65
- McKeown N, Anderson T, Balakrishnan H, Parulkar GM, Turner JS (2008) Openflow: enabling innovation in campus networks. *ACM SIGCOMM Comput Commun Rev* 38(2):69–74
- Wang K, Yin H, Quan W, Min G (2018) Enabling collaborative edge computing for software defined vehicular networks. *IEEE Netw* 32(5):112–117
- Xylomenos G, Ververidis CN, Siris VA, Fotiou N, Tsilopoulos C, Vasilakos X, Katsaros KV, Polyzos GC (2014) A survey of information-centric networking research. *IEEE Commun Surv Tutor* 16(2):1024–1049
- Zhang H, Dong P, Quan W, Hu B (2015) Promoting efficient communications for high-speed railway using smart collaborative networking. *IEEE Wirel Commun* 22(6):92–97
- Zhang H, Quan W, Chao HC, Qiao C (2016) Smart identifier network: a collaborative architecture for the future internet. *IEEE Netw* 30(3):46–51

Smart Infrastructure

- [Architecture and Data Management for Smart Community Information Platform](#)

Smart Metering, Specific Challenges, and Solution Approaches

Christian Wietfeld
Communication Networks Institute (CNI),
Faculty of Electrical Engineering and
Information Technology, TU Dortmund
University, Dortmund, Germany

Synonyms

Advanced automated meter reading (AMR);
Advanced metering infrastructure (AMI);
Intelligent metering system

Definition

Smart Metering describes an application which allows remotely determining the consumption of electricity in a household or enterprise by means of a communication network. Beyond the mere meter reading, a smart meter may also allow for the remote control of generators and consumers of electrical energy, so-called distributed energy resources (DER). Smart Metering enables a number of services related to the operation of energy systems, so-called Smart Energy Markets (e.g., dynamic pricing, demand response) and Smart Energy Grids (e.g., monitoring and control). The concept of Smart Metering also includes other utilities, such as water, gas, and district heating.

Historical Background

The introduction of electronic and networked electricity meters, which can be read remotely, was initially motivated by cost saving potential and expected process efficiency gains: a meter with connectivity enables the utility to determine the consumption at any point in time without having to send a person to the location of

the meter, which is often not accessible without prior arrangement. At the same time, information about the energy usage can be made available to the consumer in more frequent time intervals in order to raise the awareness of the energy used to eventually motivate energy-saving behavior. A Smart Meter may also enable utilities to reduce or even switch off the energy supply remotely (e.g., in case the contract with a consumer is terminated). First-generation Smart Metering systems have been implemented in various countries on different scales using different communication technologies, including wireless technologies (Lichtensteiger et al. 2010). The largest Smart Meter rollout took place from 2001 to 2009 in Italy, where over 30 million Smart Meters were connected using primarily narrowband power line communications (PLC).

With the fundamental changes in the overall architecture of the energy system in terms of energy market deregulation and increased share of renewables to reduce carbon emissions, Smart Metering has become an important enabler to implement advanced:

- **Smart Market** mechanisms (e.g., automated control of household appliances based on dynamic pricing depending on the availability of renewable energy)
- **Smart Grid** control (e.g., stability monitoring and control of the electricity grid in case of increasingly highly distributed and volatile energy production from distributed renewable energy sources)

Those second-generation Smart Metering systems require a higher communication performance in terms of data rate, response times, and capacity. Therefore a number of communication options are being discussed, including wireless networks (Wietfeld et al. 2011). In some regions the rollout of Smart Meters has been mandated by dedicated regulation, for example, in the European Union, a corresponding directive mandates that 80% of consumers shall be equipped with intelligent metering systems by 2020 (European Parliament 2012).

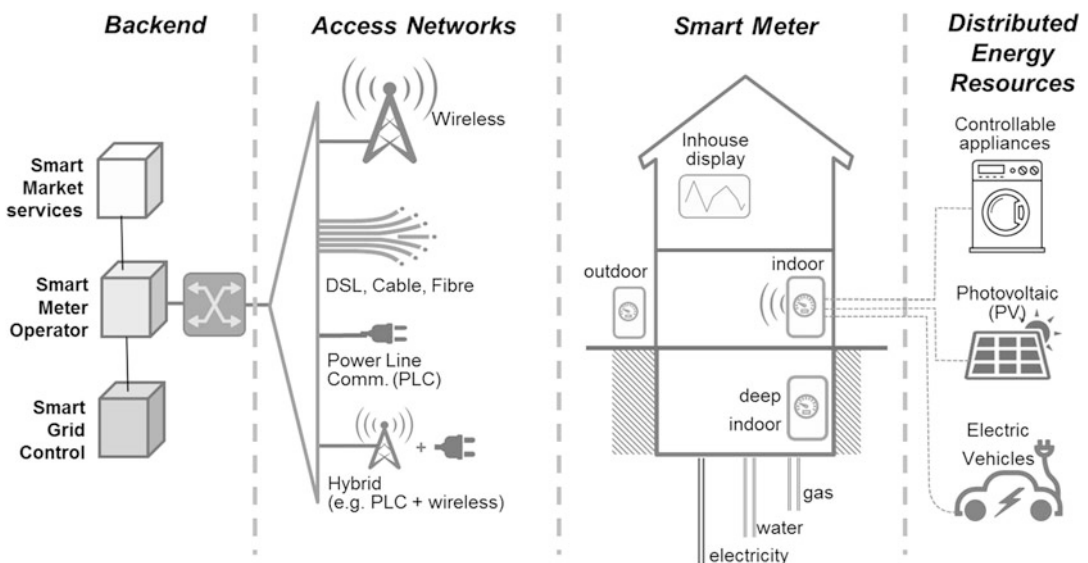
Requirements and Challenges

A Smart Metering application (see Fig. 1 for a high-level architecture overview) typically comprises the *Smart Meters* installed at the households which are mainly consumers of electrical energy but to an increasing amount may also be producers, e.g., by photovoltaic (PV) modules. The Smart Meters are highly embedded, electronic systems of a size similar to traditional meters to fit in the original installation. They account the exact amount consumed and produced energy with a given resolution. The *Smart Meter operator* collects the information in the back end and provides interfaces to Smart Market and Smart Grid control instances. The resolution of the metering information is not only defined by technical boundary conditions (e.g., amount of storage memory), but may also take into account privacy concerns of the consumers (Finster and Baumgart 2015).

A key element of Smart Meters are their *networking capabilities*: on the one hand, the metering information is delivered via wide-area communication to the accounting system of the utility to produce an accurate bill; on the other hand, the current and accumulated

metering information may also be displayed to the consumer, either via a locally connected device or via a smartphone app via the Internet. Additionally, the Smart Meter for electricity may also be connected to meters of other utilities (multi-utility) and/or distributed energy resources (DER) such as remotely controllable household appliances. Advanced Smart Meters may also monitor parameters of the energy systems, such as phase or frequency of current or voltage. Finally, the Smart Meter may even execute switching actions with DERs necessary to preserve the stability of the energy grid, e.g., by taking a PV module of the grid to prevent a critical overload or start the charging of an electric vehicle to take load off the grid.

From this wide range of potential services associated with Smart Meters, it becomes apparent that the *IT security requirements* play an important role when implementing a Smart Metering system: the coordinated attack on a large fleet of Smart Meters is considered to be a severe risk for the stability of the energy system, as energy loads may be manipulated to even physically harm the overall energy grid (McDaniel and McLaughlin 2009).



Smart Metering, Specific Challenges, and Solution Approaches, Fig. 1 High-level architecture of Smart Metering application

The *wide-area communication requirements* of Smart Meters can be described by the following parameters:

- **Location:** meters are often located in areas of the house, which are not a primary focus for providing network connectivity. Depending on country-specific preferences and tradition, meters may be located outside of buildings (e.g., in the USA) but also indoor and even in basements, requiring a so-called deep indoor coverage (e.g., in Germany and other European countries). Ideally, an existing meter can be simply replaced by a Smart Meter reusing the same installation location without further wiring, thereby creating a minimum of effort on the end consumer side.
- **Size of data packets:** there are different standardized protocols which describe the detailed message formats associated with Smart Metering services (Feuerhahn et al. 2011). While in regular operation, the size of the data packets related to metering and monitoring is small (kB range), in case of firmware updates, 10+ MB of data need to be transferred.
- **Latency:** within the distribution network, the latency requirements are in the range of seconds rather than ms and therefore less harsh than in the high-voltage network or other vertical industry areas (e.g., autonomous driving). To produce accurate time stamps, the latency of the communication link needs to be below a certain threshold in case a synchronization with network-based time sources is implemented (Mills 1991).
- **Network capacity:** following the abovementioned IT security concerns, the network needs to provide a sufficient capacity to handle encrypted data, as well as operate secure and authenticated connections (e.g., Transport Layer Security). Unlike the end user-driven communication traffic, Smart Meter communication traffic does not have an inherently stochastic nature, but can be controlled by the operator of the Smart Metering system. Through suitable scheduling, it can be avoided that different

Smart-Meter-related services are initiated in parallel and lead to overload situations within the communication network. If, for example, the Smart Meters shall deliver metering information in 15-min intervals, a corresponding system design can ensure an even spread of the transmission of the metering information within a network section in which the Smart Meters compete for the same network resources.

- **QoS guarantees:** in contrast to best effort consumer services, Smart Meter services need to be executed with guaranteed performance, for example, a system-critical firmware update necessary to fix severe security threat needs to be delivered to all Smart Meters within the shortest possible time frame. Due to security considerations, each update may need to be authenticated and executed individually, leading to a significantly higher network load than in times of regular operation. As a consequence, the wide-area communication system to support Smart Meters needs to provide effective prioritization and resource management.

The standard Internet connection of the end consumer – may it be based on Digital Subscriber Line (DSL), broadband cable, or fiber – is typically considered not to be suitable for advanced Smart Metering communication traffic, as it is not owned and controlled by the Smart Meter operator. Therefore dedicated network connectivity solutions are needed to implement Smart Meter infrastructures. The specific challenge lies in accessing the households through an appropriate access network (AN) or more specifically neighborhood area network (NAN). While first-generation automated meter reading systems have been to a large extent based on PLC infrastructures, wireless systems have been also considered an option for Smart Meters from early onwards. Reusing the power supply line for communication is attractive for utilities and especially broadband powerline communications is a viable option for Smart Metering systems. Yet electromagnetic interferences between PLC and

the power system may impact the performance of the communication link (Ahmed and Lampe 2013). Wireless networks offer the flexibility to be operated independently of the power grid and the end consumers' communication network. Still they can provide connectivity for the Smart Meter without further wiring inside the consumers building.

Solution Approaches Based on Wireless Networks

Smart Meter communication infrastructures may be realized as dedicated utility-specific wireless networks, and they may leverage public network infrastructures. In the following example solution approaches are described.

Cellular Networks

Data services of cellular networks are suited to fulfill the requirements for Smart Metering applications to different degrees:

- **Operating frequencies:** As a general rule of wireless networks, a lower frequency is associated with less attenuation and therefore better coverage. In particular when it comes to deep indoor penetration, networks operated at *subGHz frequencies* have significant advantages over networks operated at higher frequencies (Hägerling et al. 2014). While until recently GSM/GPRS/EDGE operated at 900 MHz was used to support first Smart Meter installations, LTE operated at subGHz (at 700 MHz and 800 MHz) has become a preferred wireless networking choice for the most recent generation of Smart Meters. UMTS is – due to its higher carrier frequency in the GHz range – not suited to meet the coverage expectations of Smart Metering services.
- **National roaming:** Due to varying basement construction and materials, coverage of 100% is difficult to achieve even for subGHz frequencies. A hybrid approach combining power line communication to provide connectivity to a difficult-to-reach basement

with a wireless connection at the PLC concentrator is one option. But by combining multiple cellular networks with partially complementary coverage, the overall coverage can be enhanced as well (Wäfler and Heegaard 2015). While multi-homing approaches involving multiple contracts and SIM cards are costly, international SIM cards provide national roaming capabilities and enable a flexible association with alternative networks. National roaming is usually not supported as standard feature due to regulatory constraints, yet it can be easily supported by introducing SIM cards issued in another country. Based on associated roaming agreements, the access to any available cellular network is enabled. Studies show the benefit of national roaming for critical infrastructures (Mattioli and Dekker 2013).

- **QoS guarantees:** Although prioritization options have become technically available by *IMS* (IP-based Multimedia Subsystem)-enabled policy control function, service providers do not yet provide service guarantees to third parties. With the future adoption of *network slicing technologies* (Rost et al. 2017) and software-defined networking (Dorsch et al. 2014) for public 4G and 5G cellular networks, exclusive network resources can be offered to Smart Meter operators to allow for the management of resources as if they were operating their own network.
- **Machine-type communications:** While *GPRS/EDGE machine-to-machine communications* is an option for early Smart Meter deployments with limited penetration, there are network capacity limitations in case of higher penetration rates and advanced security requirements (e.g., fast firmware upgrades requiring higher bandwidths). Therefore *LTE*-based deployments are considered to be a sustainable solution approach, addressing both capacity as well as QoS requirements. Recent additions to LTE addressing specifically *machine-type communications* eMTC (evolved machine-type communications) and NB-IoT (narrowband Internet of things) are of

interest in the context of Smart Metering, as they allow for cost-efficient modems using lower bandwidths and provide so-called *extended coverage* operation modes. But the extended coverage modes need to be carefully chosen as the improved coverage needs to be traded with lower spectrum efficiency, lower data rate, and significantly higher latency (Liberg et al. 2018).

- **Networks at 450 MHz:** While the CDMA-based UMTS technology has been identified as not ideally suited for Smart Metering due to its typical operating frequencies, the alternative third-generation mobile network technology CDMA-2000 has been adapted to **450 MHz** and has been implemented as dedicated communication network for utilities (Nedevschi et al. 2007). The focus of those CDMA-450 networks is usually on system-critical Smart Grid operations but may also include Smart Metering applications. As LTE can also be operated in a 450 MHz band, it can be expected that LTE-450 will gradually replace CDMA-450 installations as fourth-generation mobile networking technology.

Wireless Mesh Networks

While cellular wireless networks are certainly a viable option for Smart Metering applications, also alternative wireless technologies have been proposed. The drivers for those alternatives are the operation in less costly or even *license-free frequency bands* and the use of cost-efficient system technology, which spare unnecessary functionalities such as mobility support. The solution options range from short-range systems which just bridge the distance between household and the road to allow for collection of metering data by a bypassing vehicle to mesh networks which provide wide-area coverage similar to cellular networks, but with a dedicated focus.

Various *proprietary* (Lichtensteiger et al. 2010) as well as *standardized* wireless mesh technologies, for example, IEEE 802.15.4 (Fadlullah et al. 2011) and IEEE 802.11 (Böcker et al. 2017), have been proposed in recent years and implemented in various scenarios.

As the operation in unlicensed bands is typically restricted in terms of transmit power and duty cycle, those technology options have been particularly been successful in markets where outdoor coverage is sufficient. Leveraging *relay functionality* is essential for extending the coverage; at the same time, capacity bottlenecks need to be avoided by careful network planning.

For advanced Smart Metering applications supporting system-critical functionality, the operation in unlicensed bands, although it may be less costly, does often not fulfill the requirements of a critical infrastructure operation regarding QoS (Böcker et al. 2017).

Conclusions and Outlook

Smart Metering is an enabler for the automation of the energy system to support Smart Market and Smart Grid services. It is key example of a new class of vertical applications with challenging performance requirements regarding scalability and QoS guarantees. As result of region-specific requirements and boundary conditions, different wireless solution options for Smart Metering systems are viable.

Cross-References

- ▶ 5G
- ▶ Edge Caching
- ▶ IEEE 802.11
- ▶ LTE: Long-Term Evolution
- ▶ NB-IoT
- ▶ Network Slicing-Enabled Green C-RAN
- ▶ Resource Management
- ▶ Scheduling
- ▶ Wireless Fading Channel Model for 5G and Future

References

- Ahmed MO, Lampe L (2013) Power line communications for low-voltage power grid tomography. IEEE Trans Commun 61(12):5163–5175

- Böcker S, Jörke P, Wietfeld C (2017) Performance evaluation of an IEEE 802. 11 mesh-based smart market and smart grid communication system. In: IEEE international conference on smart grid communications, pp 183–188
- Dorsch N, Kurtz F, Georg H et al (2014) Software-defined networking for Smart Grid communications: applications, challenges and advantages. In: 2014 IEEE international conference on smart grid communications, SmartGridComm 2014
- European Parliament (2012) Directive 2012/27/EU of the European Parliament and of the council of 25 October 2012 on energy efficiency. Off J Eur Union Dir 1–56. https://doi.org/10.3000/19770677.L_2012.315.eng
- Fadlullah ZM, Fouda MM, Kato N et al (2011) Toward intelligent machine-to-machine communications in smart grid. IEEE Commun Mag 49:60
- Feuerhahn S, Zillgith M, Wittwer C, Wietfeld C (2011) Comparison of the communication protocols DLMS/COSEM, SML and IEC 61850 for smart metering applications. In: 2011 IEEE international conference on smart grid communications, SmartGridComm 2011
- Finster S, Baumgart I (2015) Privacy-aware smart metering: a survey. IEEE Commun Surv Tutorials 16: 1732
- Hägerling C, Ide C, Wietfeld C (2014) Coverage and capacity analysis of wireless M2M technologies for smart distribution grid services. In: 2014 IEEE international conference on smart grid communications, SmartGridComm 2014
- Liberg O, Sundberg M, Wang Y-PE, et al (2018) The cellular internet of things. In: Cellular internet of things: Technologies, Standards, and Performance, Academic Press
- Lichtensteiger B, Bjelajac B, Mueller C, Wietfeld C (2010) RF mesh systems for smart metering: system architecture and performance. In: 2010 first IEEE international conference on smart grid communications, pp 379–384
- Mattioli R, Dekker M (2013) National roaming for resilience. ENISA European Union Agency for Network and Information Security
- McDaniel P, McLaughlin S (2009) Security and privacy challenges in the smart grid. IEEE Secur Priv 7:75
- Mills DL (1991) Internet time synchronization: the network time protocol. IEEE Trans Commun 39(10):1482–1493
- Nedevschi S, Surana S, Du B et al (2007) Potential of CDMA450 for rural network connectivity. IEEE Commun Mag 45:128
- Rost P, Mannweiler C, DiS M et al (2017) Network slicing to enable scalability and flexibility in 5G mobile networks. IEEE Commun Mag 55:72
- Wäfler J, Heegaard PE (2015) How to use mobile communication in critical infrastructures: a dependability analysis. In: Koornneef F, van Gulijk C (eds) Computer safety, reliability, and security. Springer International Publishing, Cham, pp 335–344
- Wietfeld C, Georg H, Gröning S et al (2011) Wireless M2M communication networks for smart grid applications. In: 17th European wireless conference 2011, EW 2011

Smartphone Security and Privacy

► Mobile Security and Privacy

SMDP

► Handoff Management in Wireless Communication-Based Train Control Systems

Social Cooperation System

► Social IoT Crowd-Sourcing on Disaster Reduction

Social Data Management

► Architecture and Data Management for Smart Community Information Platform

Social Group Utility Maximization

Xu Chen

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

Synonyms

Social group utility maximization

Definition

Social group utility maximization (SGUM) for cooperative wireless networking takes into account both social relationships and physical coupling among users. Specifically, instead of maximizing its individual utility or the overall network utility, each user aims to maximize its social group utility that hinges heavily on its social tie structure with other users. This framework provides rich modeling flexibility and spans the continuum between noncooperative game and network utility maximization (NUM) – two traditionally disjoint paradigms for network optimization.

Historical Background

Mobile networks have been projected to continue growing rapidly in the foreseeable future. Different from other networks (e.g., sensor networks), a distinctive characteristic of mobile networks is that mobile devices are carried and operated by human beings. With the explosive growth of online social networks such as Facebook and Twitter, more and more people are actively involved in online social interactions, and social relationships among people are hence extensively broadened and significantly enhanced. Then it is natural to ask whether “social relationships” among mobile users can impact the communications and interactions among their devices, and if yes, how can the

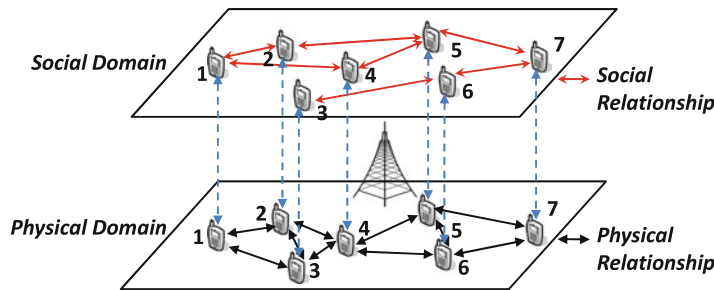
social structure be cleverly leveraged? With this insight, as illustrated in Fig. 1, a mobile network is viewed as an overlay/underlay system where a “virtual social network” overlays the physical communication network (the “social network” is virtual, in the sense that the social tie structure therein results from existing human relationship and online social networks) and leverages the intrinsic social tie structure among mobile users to facilitate cooperative wireless networking.

With this motivation, a novel *social group utility maximization* (SGUM) framework is advocated to take into account both the users’ social relationships and physical coupling. As illustrated in Fig. 1, a key observation is that users are coupled not only in the physical domain due to the physical relationship (e.g., interference) but also in the social domain due to the social ties among them. It would be a win-win case for users to help those users having social ties with them. Specifically, the distributed decision-making problem for cooperative networking among users is casted as a SGUM game, where each user maximizes its social group utility, defined as the sum of its own individual utility and the weighted sum of the utilities of other users having social tie with it (Chen et al. 2014, 2015, 2016).

SGUM Framework

Physical Network Graph Model

A set of wireless users $\mathcal{N} = \{1, 2, \dots, N\}$ is considered where N is the total number of users.



Social Group Utility Maximization, Fig. 1 An illustration of the social group utility maximization framework. In the physical domain, users have different physical cou-

pling subject to physical relationships (e.g., interference). In the social domain, users have heterogeneous social coupling due to the social ties among users

The set of feasible strategies for each user $n \in \mathcal{N}$ is denoted as $\mathcal{X}_n$. For instance, a strategy $x \in \mathcal{X}_n$ can be choosing either a channel or a power level for wireless transmissions. Subject to heterogeneous physical constraints, the strategy set $\mathcal{X}_n$ can be user-specific. For example, the strategy set $\mathcal{X}_n$ can be a set of feasible relay users that are in vicinity of user n for cooperative communication.

To capture the physical coupling among the users in the physical domain, the *physical graph* $\mathcal{G}^p = \{\mathcal{N}, \mathcal{E}^p\}$ is introduced (see, e.g., Fig. 1). Here the set of users $\mathcal{N}$ is the vertex set, and $\mathcal{E}^p \triangleq \{(n, m) : e_{nm}^p = 1, \forall n, m \in \mathcal{N}\}$ is the edge set where $e_{nm}^p = 1$ if and only if users n and m have physical coupling, i.e., users n and m can affect each other's payoff by taking some actions. For example, two users have the physical coupling if they can cause interference to each other when using the same channel for data transmission. Then the set of users that have physical coupling with user n is also denoted as $\mathcal{N}_n^p \triangleq \{m \in \mathcal{N} : e_{nm}^p = 1\}$.

Let $\mathbf{x} = (x_1, \dots, x_N) \in \prod_{n=1}^N \mathcal{X}_n$ be the strategy profile of all users. Given the strategy profile $\mathbf{x}$, the individual utility function of user n is denoted as $U_n(\mathbf{x})$, which represents the payoff of user n , accounting for the physical coupling among users. For example, $U_n(\mathbf{x})$ can be the achieved data rate or the satisfaction of quality of service (QoS) requirement of user n under the strategy profile $\mathbf{x}$. Note that in general, the underlying physical graph plays a critical role in determining the individual utility $U_n(\mathbf{x})$. For example, users' achieved data rates are determined by the interference graph and channel quality.

Social Network Graph Model

The social graph model is next introduced to describe the social ties among users. The underlying rationale of considering social ties is that the handheld devices are carried by human beings and the knowledge of human social ties can be utilized to enhance the performance of cooperative networking.

Specifically, the *social graph* $\mathcal{G}^s = \{\mathcal{N}, \mathcal{E}^s\}$ is introduced to model the social ties among the users. Here the vertex set is the same as

the user set $\mathcal{N}$, and the edge set is given as $\mathcal{E}^s = \{(n, m) : e_{nm}^s = 1, \forall n, m \in \mathcal{N}\}$ where $e_{nm}^s = 1$ if and only if users n and m have social tie between each other, which can be kinship, friendship, or colleague relationship between two users. Furthermore, for a pair of users n and m who have a social edge between them on the social graph, the strength of social tie is formalized as $w_{nm} \in [0, 1]$, with a higher value of w_{nm} being a stronger social tie. User n 's *social group* $\mathcal{N}_n^s$ is defined as the set of users that have social ties with user n , i.e., $\mathcal{N}_n^s = \{m : e_{nm}^s = 1, \forall m \in \mathcal{N}\}$.

Based on the physical and social graph models above, users are coupled in the physical domain due to the physical relationships and also coupled in the social domain due to the social ties among them. It would be a win-win case for users to help those users having social ties with them. With this insight, the *social group utility* of each user n is then defined as

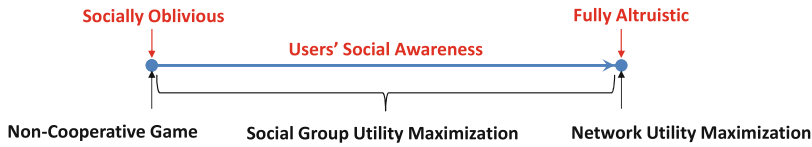
$$S_n(\mathbf{x}) = U_n(\mathbf{x}) + \sum_{m \in \mathcal{N}_n^s} w_{nm} U_m(\mathbf{x}). \quad (1)$$

It follows that the social group utility of each user consists of two parts: (1) its own individual utility and (2) the weighted sum of the individual utilities of other users having social ties with it. In a nutshell, the social group utility function captures the feature that each user is socially aware and cares about the users having social ties with it.

Social Group Utility Maximization Game

The distributed decision-making problem among the users is then considered, aiming to maximize their social group utilities. Let $\mathbf{x}_{-n} = (x_1, \dots, x_{n-1}, x_{n+1}, \dots, x_N)$ be the set of strategies chosen by all other users except user n . Given the other users' strategies $\mathbf{x}_{-n}$, user n wants to choose a strategy $x_n \in \mathcal{X}_n$ to maximize its social group utility, i.e.:

$$\max_{x_n \in \mathcal{X}_n} S_n(x_n, \mathbf{x}_{-n}), \forall n \in \mathcal{N}.$$



Social Group Utility Maximization, Fig. 2 Network optimization with different users' social awareness levels

The distributed nature of the problem above naturally leads to a formulation based on game theory such that each user aims to maximize its social group utility. The decision-making problem among the users is thus formulated as a strategic game $\Gamma = (\mathcal{N}, \{\mathcal{X}_n\}_{n \in \mathcal{N}}, \{S_n\}_{n \in \mathcal{N}})$, where the set of users $\mathcal{N}$ is the set of players, $\mathcal{X}_n$ is the set of strategies for each user n , and the social group utility function S_n of each user n is the payoff function of player n . The game Γ is called as the SGUM game, and the concept of *socially aware Nash equilibrium* (SNE) is defined as follows:

Definition 1 A strategy profile $\mathbf{x}^* = (x_1^*, \dots, x_N^*)$ is a socially aware Nash equilibrium of the SGUM game if no player can improve its social group utility by unilaterally changing its strategy, i.e.:

$$x_n^* = \arg \max_{x_n \in \mathcal{X}_n} S_n(x_n, x_{-n}), \forall n \in \mathcal{N}.$$

It is worth noting that under different social graphs, the proposed SGUM game formulation can provide rich flexibility for modeling the network optimization problem (as illustrated in Fig. 2). When the social graph consists of isolated nodes with $w_{nm} = 0$ for any $n, m \in \mathcal{N}$ (i.e., all users are selfish), the SGUM game degenerates to the noncooperative game formulation. When the social graph is fully meshed with edge weight $w_{nm} = 1$ for any $n, m \in \mathcal{N}$ (i.e., each user is fully altruistic and cares enough about other users), the SGUM game becomes the network utility maximization problem, which aims to maximize the system-wide utility. The SGUM framework in this study is applicable to general social graphs and hence can bridge the

gap between noncooperative game and network utility maximization – two traditionally disjoint paradigms for network optimization.

Due to its rich modeling flexibility, the SGUM framework has been widely applied in many networking issue studies, such as power control, random access control, spectrum access, computation offloading, data caching, and wireless jamming (Chen et al. 2016; Gong et al. 2014, 2017, 2015; Tang et al. 2016; Zhu et al. 2017).

Cross-References

► [Efficiency and Pareto Optimality](#)

References

- Chen X, Gong X, Yang L, Zhang J (2014) A social group utility maximization framework with applications in database assisted spectrum access. In: IEEE INFOCOM, pp 1959–1967
- Chen X, Proulx B, Gong X, Zhang J (2015) Exploiting social ties for cooperative d2d communications: a mobile social networking case. *IEEE/ACM Trans Netw* 23(5):1471–1484
- Chen X, Gong X, Yang L, Zhang J (2016) Exploiting social tie structure for cooperative wireless networking: a social group utility maximization framework. *IEEE/ACM Trans Netw* 24(6):3593–3606
- Gong X, Chen X, Zhang J (2014) Social group utility maximization in mobile networks: from altruistic to malicious behavior. In: 2014 48th annual conference on information sciences and systems (CISS). IEEE, pp 1–6
- Gong X, Duan L, Chen X (2015) When network effect meets congestion effect: leveraging social services for wireless services. In: Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing, pp 147–156

- Gong X, Chen X, Xing K, Shin DH, Zhang M, Zhang J (2017) From social group utility maximization to personalized location privacy in mobile networks. *IEEE/ACM Trans Netw* 25(3):1703–1716
- Tang L, Chen X, He S (2016) When social network meets mobile cloud: a social group utility approach for optimizing computation offloading in cloudlet. *IEEE Access* 4:5868–5879
- Zhu K, Zhi W, Chen X, Zhang L (2017) Socially motivated data caching in ultra-dense small cell networks. *IEEE Netw* 31(4):42–48

Social Internet of Things

Han-Chieh Chao^{1,2}, Hsin-Hung Cho², Wei-Che Chien³, Shih-Yun Huang¹, and Ting-Mei Li¹

¹Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan

²Department of Computer Science and Information Engineering, National Ilan University, Yilan, Taiwan

³Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan, Republic of China

Synonyms

Social network service and the Internet of things

Definition

The Social Internet of Things is a social network that includes the Internet of Things, allowing you to share services with others through social networks. Each Internet of Things (IoT) contains a number of technologies, including being able to search each other's Internet resources through social networks and work with your online neighbors to achieve common goals.

Historical Background

Social Network

The structure of a social network (Scott 2017) is premised on a network structure (Ray and

Bill 2003) formed by the relationships between many nodes. People or organizations are like nodes, and social relations represent the connections between these nodes. There are social interactions between people or people and organizations, and these interactions create connections that link people or organizations together to form social networks. The development of social networks started from the rise of the e-mail period, and with the progress of the Internet, social networks have gradually transformed into other areas: the family site period, instant messaging period, dating period, and the blog period, to the current period of social network services (Danah and Nicole 2007). Social networks create extremely virtual and realistic human interactions on the Internet.

Internet of Things

The Internet of Things (IoT) (Jayavardhana et al. 2013) aims to enable all types of communication on the Internet. All ordinary objects that can perform independent functions can be connected through the network. Connections are made device to device through Bluetooth, ZigBee, RFID, and other information sensing technology (Lionel et al. 2004; Jin-Shyan et al. 2007), enabling all devices to connect to the network. Through communication and information exchanges, all devices become conscious and intelligent. As the number of smart devices increases, so too does the amount of data moving between devices. In order to manage the installation and use the information more effectively, the Social Internet of Things (SIOT) (Luigi et al. 2012) is generated. By combining many social networking groups, SIOT is able to combine a large amount of data to provide more accurate answers to complex questions. And because SIOT uses social connections in social networks, the IoT can build larger social relationships with more devices. SIOT offers a wider range of smarter applications than IoT does (Luigi et al. 2014).

There are five relationship types in SIOT (Luigi et al. 2014): parental object relationship (POR), co-location object relationship (C-LOR), co-work object relationship (C-WOR), ownership

Social Internet of Things, Table 1 The architecture schematic of SIoT

Relationship type	Relationship rule	Communication type
Parental object relationship (POR)	There is a connection between the same devices made by the same manufacturer	M2M, D2D
Co-location object relationship (C-LOR)	The connection between people or devices in the same place	M2M, H2M, M2H, H2H, D2D
Co-work object relationship (C-WOR)	A collaboration object that is created when the same kind of IOT application is coshared	M2M, H2M, M2H, D2D
Ownership object relationship (OOR)	Same appliance with multiple users	H2M, M2H
Social object relationship (SOR)	The social behavior between people in life generates connections on the device	H2H, M2M, D2D

object relationship (OR), and SOR (SOR). The five relationship types are described in Table 1.

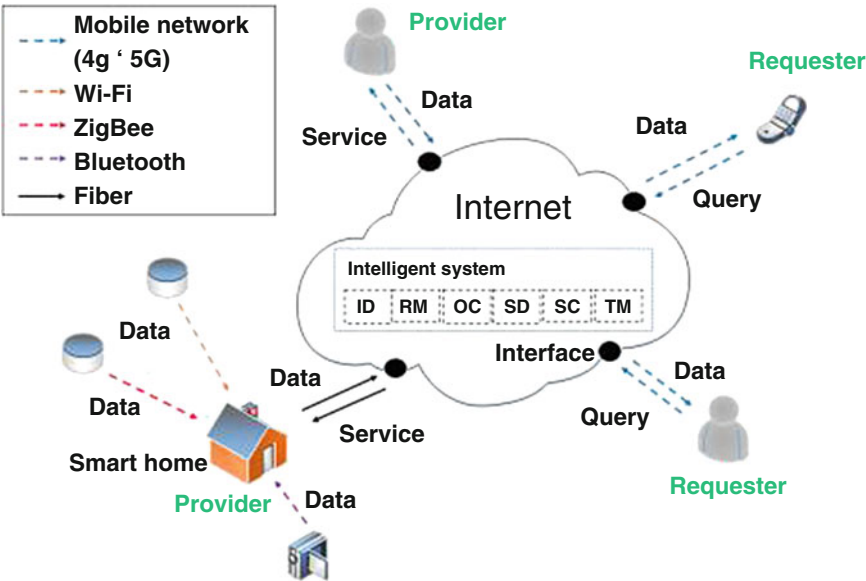
Four different communication types are used in SIoT's five relationship types: human to human (H2H), machine to machine (M2M), human to machine (H2M), and machine to human (M2H). H2H refers to a person's communication behavior after establishing a social relationship. M2M refers to devices that can communicate directly with each other over a wired or wireless network. In the M2M there is more developed D2D technology. D2D (Ometov et al. 2016) refers to direct communication between two mobile communication devices. H2M signifies communication between a person and a device. The device is an auxiliary tool controlled by the operator. M2H means that the device sends a notification to a person when an event on the device triggers a conditional alarm.

Architecture of SIoT

The components of SIoT can be divided into four parts (Ali 2015). The first part is the actor which includes humans and IoT devices. Both of these can manage data by sending data and receiving control information. The second part is an intelligent system, which is used to manage and consort interactions between different actors. The third component of SIoT is the interface; its main function is to provide an interface which enables the actor to input and query data. The fourth component is the Internet that can connect actors anywhere at any time. Finally, the communication medium is also an important part for SIoT, so that the service of a smart device can be offered to users through Wi-Fi (Huang et al. 2017), mobile communication (Al-Turjman 2017), and so on. Among these components, the intelligence system is a key technology, which can build relationships between each actor, provide suitable access permission for different groups, and discover the relationships between different groups. Atzori et al. (2012) proposed the structure of an intelligence system which included identity management, owner control (OC), relationship management (RM), service discovery (SD), service composition (SC), and trustworthiness management (TM). The architecture schematic of SIoT is shown in Fig. 1.

Grouping

Among these components, SD and SC are the first and most critical steps to build relationships among different actors and to compose them together. Many technologies are used to discover the relationship between actors by enhancing semantic reasoning for greater context awareness (Maarala et al. 2017). Shen et al. (2017) adopt the semantics of a top-k query to search the optimal SIoT group objects. Kang et al. (2016) proposed a grouping method based on social correlations and the consistence of the groups according to similar node information. Gil et al. (2016) summarized the context aware method of SIoT, which includes information retrieval (IR) (Rau et al. 2015), infor-



Social Internet of Things, Fig. 1 The architecture schematic of SIoT

Social Internet of Things, Table 2 Analysis of communication type of SIoT works

Work	Key communication types				
	M2M	H2M	M2H	H2H	D2D
Shen et al. (2017)	V				
Kang et al. (2016)				V	
Rau et al. (2015)				V	
Zhao et al. (2015)	V				
Romer et al. (2010)		V	V		
Al-Turjman (2017)		V	V		
Huang et al. (2017)	V				
Ometov et al. (2016)					V

information extraction (Zhao et al. 2015), and a search engine for the Web of Things (Romer et al. 2010). It can be seen that the semantic reasoning is an important technology for SIoT, which facilitates data integration and interoperability in SIoT applications (Barnaghi et al. 2012). Table 2 is an analysis of communication types of SIoT works.

Key Applications

In order to satisfy and realize the demand for people and things to exchange their data or infor-

mation in SIoT, the deployment of IoT devices, through the information platform to exchange their information or integrate the data and using Web service or API, is required. But, in the application of SIoT, the following details must be noticed (Victoria et al. 2014): (1) the users can know their personal information and can update their status or delete it; (2) when users want some information they find interesting, they can get this information from IoT devices; (3) the users have the ability to change the status of the things/devices, e.g., using the smartphone to control the lights in their home; and (4) finally, the application should provide a trigger rule, so users have the ability to easily set the rule. Due to the fact that the goal of SIoT is to facilitate the relationships between people and things, and it may help to make the IoT smarter, there is much relevant literature and application services focus on it (Table 3).

In regard to agriculture, Abdullah and William (2016) suggest using the IoT devices to help farmers collect data; through these services, they can then share information with other farms. In the smart home (Dina et al. 2013), the IoT device deployments are targeted at household appliances and can gather information, like the temperature

Social Internet of Things, Table 3 The application of SIoT

	Application	Literature	Communication type	Common relationship type
Consumer	Smart home	Dina et al. (2013)	H2M, M2M, D2D	C-LOR, OOR, SOR
	Smart shopping	Ugo et al. (2011)	M2H, H2H	OOR, SOR
	e-health	Chaitali et al. (2016)	H2H, D2D	POR, SOR
	Gastronomy/tourist	Luca et al. (2013)	M2H, H2H	OOR, SOR
Enterprise	Infrastructure monitoring	Anderson et al. (2017)	M2H	POR, C-WOR, SOR
	Transportation	Augusto et al. (2018)	M2M, D2D	C-LOR, C-WOR, SOR
	Education	Anmol and Pradeep (2016)	H2M	C-LOR, OOR, SOR
	Agricultural	Abdullah and William (2016)	M2H, D2D	C-LOR, OOR, SOR

in a room, or it serve as an alarm for emergency situations. People can use the SIoT service to shop for interesting things (Ugo et al. 2011) and find the information required with their smart-phone. Of course, it can also be used to help the tourism industry (Luca et al. 2013). Finally, these application services can be combined with others, like transportation and infrastructure monitoring, and become part of a smart city (Panagiotis et al. 2013). Anmol and Pradeep (2016) proposed its use in the school environment to help teaching and learning. It may also be used to enhance the relationships between people and things. On the other hand (Luigi et al. 2014), since smart shopping changes the habits of consumers and markets, many enterprises have been employing this application service and specifically designing products to service the consumer.

efficiency of service development in an OS is a significant challenge in SIoT (Gubbi et al. 2013). In the SIoT, IoT devices include different service types; however, the service working time is based on the capacity of the battery. Highly complex interactions will make these IoT devices run out of energy quickly. (4) **Security:** The security challenges that include privacy, trustworthiness, and integrity of data are at risk in three ways (Eagle et al. 2006). The first is third-party access to data; the second is collector use and distribution of data; and the third is when data are mixed with other data. SIoT aims to build relationships between friends in order to reduce the communication cost. The security issues of data security, data theft protection, and network attacks among different SIoT groups remain challenges that SIoT must overcome.

Challenge of SIoT

There are some derivative challenges intrinsic to the development of SIoT. (1) **Compatibility:** Seeking the compatibility of heterogeneous devices from different vendors, a standard operating system (OS) must necessarily be developed because certain applications need to interoperate with each other (Ahmed et al. 2016). (2) **Accuracy and Reliability:** For sensitive SIoT applications, such as vehicle to vehicle, ensuring accuracy and reliability is a challenge for SIoT. Thus, a highly precise system is needed (Afzal et al. 2017). (3) **Energy Efficiency:** The energy

Cross-References

- Internet of Things
- Internet of Things: Architecture, Key Applications, and Security Impacts
- Mobile Social Network

References

Abdullah N, William I (2016) Developing a human-centric agricultural model in the IoT environment. In: International conference on Internet of Things and Applications (IOTA), Pune, Jan 2016

- Afzal B et al (2017) Enabling IoT platforms for social IoT applications: vision, feature mapping, and challenges. *Futur Gener Comput Syst*. <https://doi.org/10.1109/TDSC.2015.2420552>
- Ahmed E et al (2016) Internet-of-Things-based smart environments: state of the art, taxonomy, and open research challenges. *IEEE Wirel Commun* 23(5):10–16
- Ali DH (2015) A social Internet of Things application architecture: applying semantic web technologies for achieving interoperability and automation between the cyber, physical and social worlds. Ph.D. thesis, Institute National des Telecommunications
- Al-Turjman F (2017) 5G-enabled devices and smart-spaces in social-IoT: an overview. *Futur Gener Comput Syst*. <https://doi.org/10.1016/j.future.2017.11.035>
- Anderson A et al (2017) Reliability analysis of an IoT-based smart parking application for smart cities. In: IEEE international conference on big data (Big Data), Boston, MA, USA, 11–14 Dec 2017
- Anmol S, Pradeep Y (2016) Augmenting Tutoring of students using tangible smart learning objects. In: International conference on Internet of Things and Applications (IOTA), Pune, Jan 2016
- Atzori L et al (2012) The social internet of things (SIOT)—when social networks meet the internet of things: concept, architecture and network characterization. *Comput Netw* 56(16):3594–3608
- Augusto JVN et al (2018) Fog-based crime-assistance in smart IoT transportation system. *IEEE Access* 6:11101–11111
- Barnaghi P et al (2012) Semantics for the internet of things: early progress and back to the future. *Int J Semant Web Inf Syst* 8(1):1–21
- Chaitali K et al (2016) Health companion device using IoT and wearable computing. In: International conference on Internet of Things and Applications (IOTA), Pune, Jan 2016
- Danah MB, Nicole BE (2007) Social network sites: definition, history, and scholarship. *J Comput-Mediat Commun* 13(1):210–230
- Dina H et al (2013) A framework for social device networking. In: International conference on IEEE distributed computing in sensor systems (DCOSS), Cambridge, May 2013
- Eagle N et al (2006) Reality mining: sensing complex social systems. *Pers Ubiquit Comput* 10(4):255–268
- Gil D et al (2016) Internet of things: a review of surveys based on context aware intelligent services. *Sensors* 16(7):1069
- Gubbi J et al (2013) Internet of things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comp Syst* 29(7):1645–1660
- Huang CM et al (2017) The social internet of thing (S-IOT)-based Mobile group handoff architecture and schemes for proximity service. *IEEE Trans Emerg Top Comput* 5(3):425–437
- Jayavardhana G, Rajkumar B, Slaven M et al (2013) Internet of things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comput Syst* 29(7):1645–1660
- Jin-Shyan L, Yu-Wei S, Chung-Chou S (2007) A comparative study of wireless protocols: bluetooth, UWB, ZigBee, and Wi-Fi. In: 33rd annual conference of the IEEE Industrial Electronics Society, Taipei, Taiwan, p 46
- Kang D H et al (2016) Social correlation group generation mechanism in social IoT environment. In: 2016 eighth international conference on ubiquitous and future networks (ICUFN), Vienna, Austria, July 2016, pp 514–519
- Lionel MN, Yunhao L, Yiu Cho L et al (2004) LAND-MARC: indoor location sensing using active RFID. *Wirel Netw* 10(6):701–710
- Luca C et al (2013) Interacting with social networks of intelligent things and people in the world of gastronomy. *ACM Trans Intell Syst Technol* 3(1):4
- Luigi A, Antonio I, Giacomo M et al (2012) The social internet of things (siot)—when social networks meet the internet of things: concept, architecture and network characterization. *Comput Netw* 56(16):3594–3608
- Luigi A, Antonio I, Giacomo M (2014) From “smart objects” to “social objects”: the next evolutionary step of the internet of things. *IEEE Commun Mag* 52(1):97–105
- Maarala AI et al (2017) Semantic reasoning for context-aware internet of things applications. *IEEE Internet Things* 4(2):461–473
- Ometov A et al (2016) Toward trusted, social-aware D2D connectivity: bridging across the technology and sociality realms. *IEEE Wirel Commun* 23(4):103–111
- Panagiotis V et al (2013) Enabling smart cities through a cognitive management framework for the internet of things. *IEEE Commun Mag* 51(6):102–111
- Rau PLP et al (2015) Exploring interactive style and user experience design for social web of things of Chinese users: a case study in Beijing. *Int J Hum-Comput Stud* 80:24–35
- Ray R, Bill M (2003) Network structure and knowledge transfer: the effects of cohesion and range. *Adm Sci Q* 48(2):240–267
- Romer K et al (2010) Real-time search for real-world entities: a survey. *Proc IEEE* 98(11):1887–1902
- Scott J (2017) Social network analysis. Sage, Los Angeles
- Shen CY et al (2017) Task-optimized group search for social Internet of Things. In: EDBT, Venice, Italy, Nov 2017, pp 108–119
- Ugo BC et al (2011) Design and development of a social shopping experience in the IoT domain: the ShopLovers solution. In: 19th international conference on software, telecommunications and computer networks (SoftCOM), Split, Sept 2011. IEEE
- Victoria B et al (2014) From “smart objects” to “social objects”: the next evolutionary step of the Internet of Things. *IEEE Commun Mag* 52(1):97–105
- Zhao F et al (2015) Topic-centric and semantic-aware retrieval system for internet of things. *Inf Fusion* 23:33–42

Social Internet of Things (SIoT)

► Applications and Architectures of Social Internet of Things

Social IoT Crowd-Sourcing on Disaster Reduction

Chih-Hao Liu and Chia-Feng Chiang
Information Division, National Science and
Technology Center for Disaster Reduction, New
Taipei City, Taiwan

Synonyms

Social cooperation system

Definition

Social IoT crowd-sourcing regulates users on the social network that used to publish useful information to cooperate a common task.

Historical Background

Social Internet of Things (SIoT) is a special type of IoT which provides users with fast and convenient information dissemination based on the connection of social network environment (Serrat; Atzori et al. 2012). Various kinds of data can spread rapidly through the social network (Houston et al. 2014). Users can share information as a post based on what they see and hear. When a disaster happened, people in the disaster area usually use a mobile phone to publish the disaster information with the picture to the social network. When other people see the disaster information, they usually leave a comment or upload the new information to the social network (Houston et al. 2014). The people

who proactive publish the disaster information on the social network, we called that is citizen cooperation. The citizen notification is the useful to gather the disaster information during disaster response state. And a tool with map is necessary to recruit people to gather disaster information.

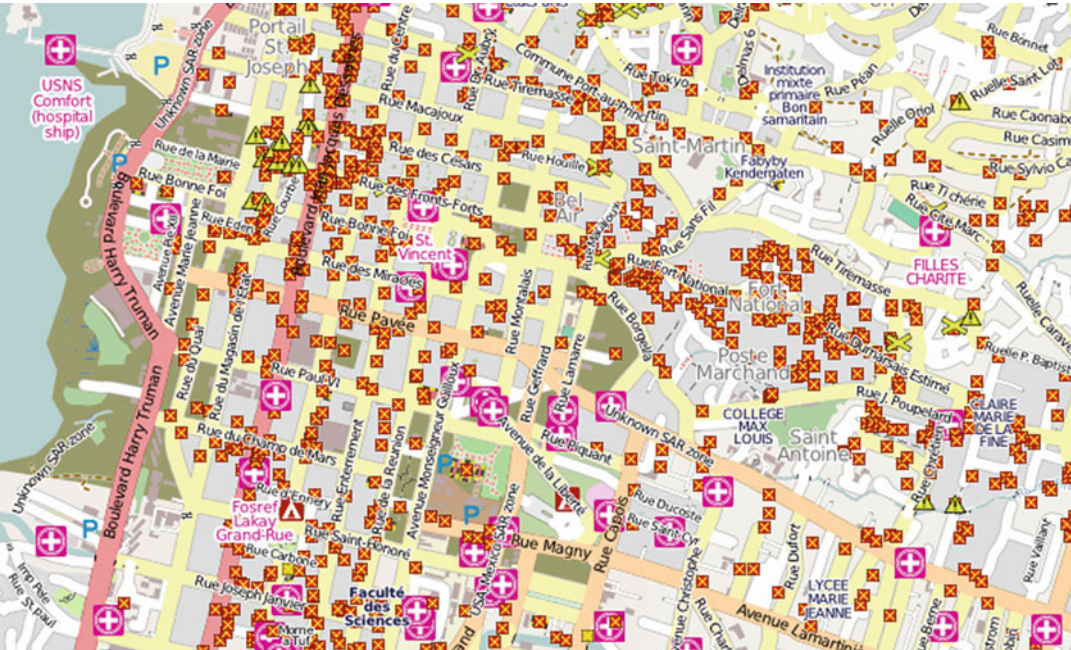
The Crowd-Sourcing System (Prpića et al. 2015; Zhao and Zhu 2014) is a mode of information acquisition. Some unspecified users on the Internet collaborate on a certain task to provide information to meet the common task (Karger et al. 2014). Each user only contributes a small amount of effort (time, information, etc.), and the entire system will gather a large amount of information. This mode can provide diversity, immediacy, and scale information (Xu et al. 2016; Lina et al. 2018). However, this approach requires many spontaneous users. Thus, we propose the citizen cooperation map to recruit users and gather the disaster information.

Disaster Reduction Application

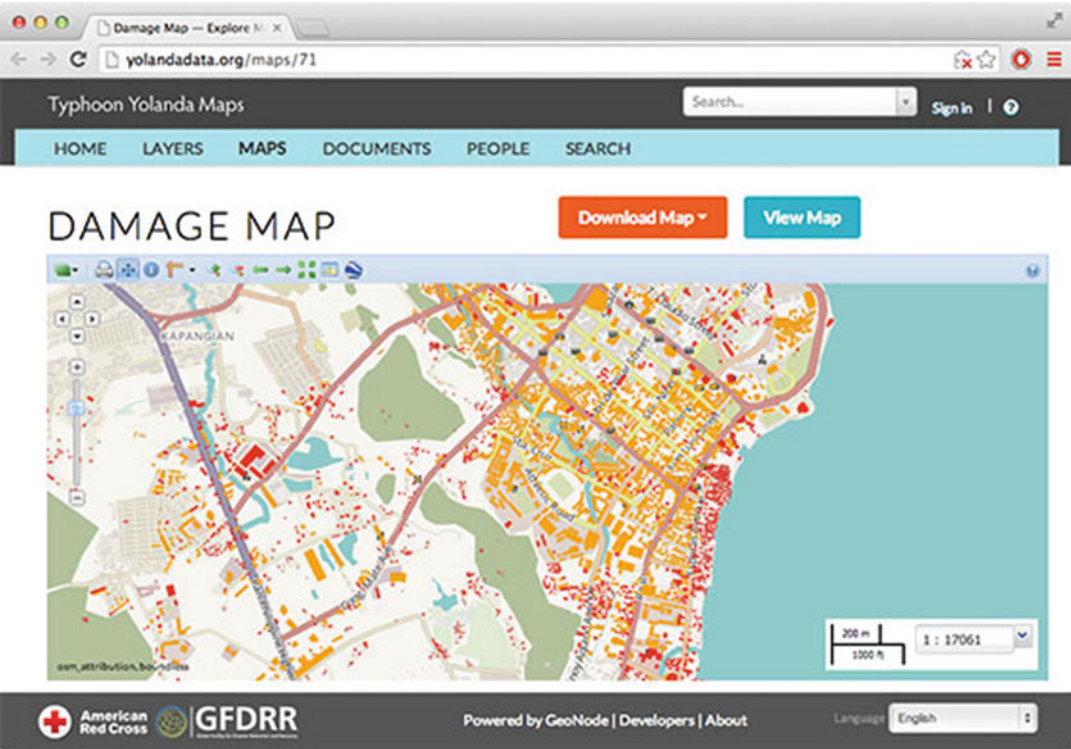
The 2010 Haiti earthquake was a magnitude 7.0 earthquake (2010 Haiti earthquake). The government has no a GIS system to integrate the disaster information. And the foundation map has not been updated for 20 years. The OSM's Humanitarian OpenStreetMap Team (HOT) spent seven days conducting field surveys and mapping by volunteers on the Internet. The map (Fig. 1) completed in a week has become an important basis for disaster relief (2010 Haiti earthquake damage map).

Many people in the disaster area use Twitter to exchange information because of the lack of telephone communications. They publish first-hand disaster information through social networks. In addition, users all over the world donate materials through social websites such as Twitter and Facebook.

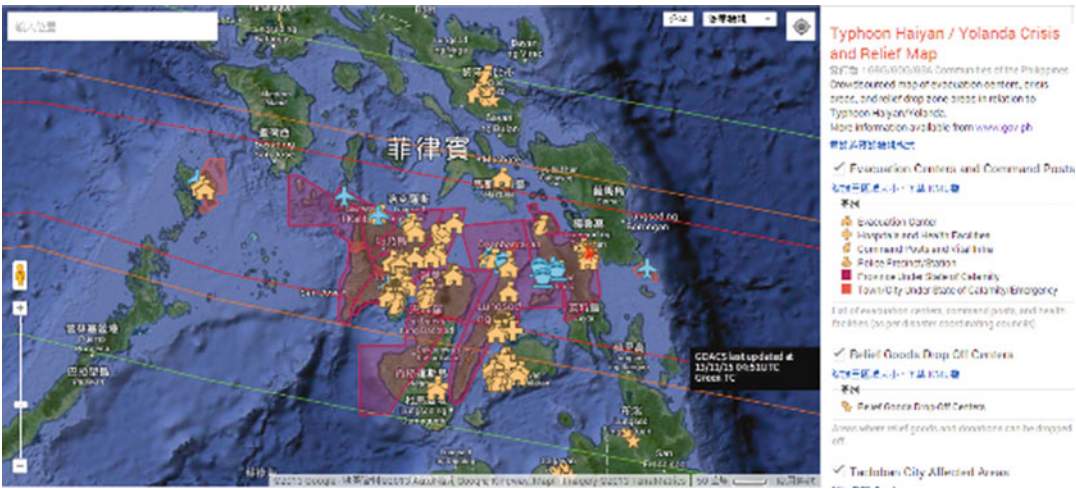
The typhoon Haiyan invades Philippines on 2013 (Typhoon Haiyan). The International Red Cross humanitarian aid organization immediately rushed to the disaster area to assist with disaster relief and recovery work, but encountered the plight of the lack of local maps. Launched by the



Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 1 The disaster map of Haiti earthquake



Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 2 The basic map of Typhoon Haiyan



Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 3 The crisis map of Typhoon Haiyan

Humanitarian OpenStreetMap Team and assisted by 410 international OSM volunteers, the basic maps (Fig. 2) of the disaster area were quickly established to provide on-site rescue organization personnel (Typhoon Yolanda Map).

During the Typhoon Haiyan, the GDG Philippines, GBG Philippines, and other nongovernmental community groups in the Philippines jointly issued a Google Crisis map (Fig. 3) to annotate disaster warning areas and material distribution areas (Typhoon Haiyan/Yolanda Crisis and Relief Map). The two examples in the above are mainly initiated by professional organizations, and the relief agencies are responsible for integrating first-hand information on the disaster.

The Typhoon Nepartak landed at Taimali Township, Taitung County, around 5:50 on July 8 and left Taiwan at 14:30 (Typhoon Nepartak 2016). The actual time to influence Taiwan was July 8. The administrator of the group of Facebook-台東臉書災害通報 used the group as a temporary disaster collection center and called on members to actively provide the latest disaster information. One of the members even establishes a collaboration map to let users upload disaster information (Fig. 4).

Base on the concept of social Internet of Things, the citizen cooperation map (Fig. 5) is established by the National Science and

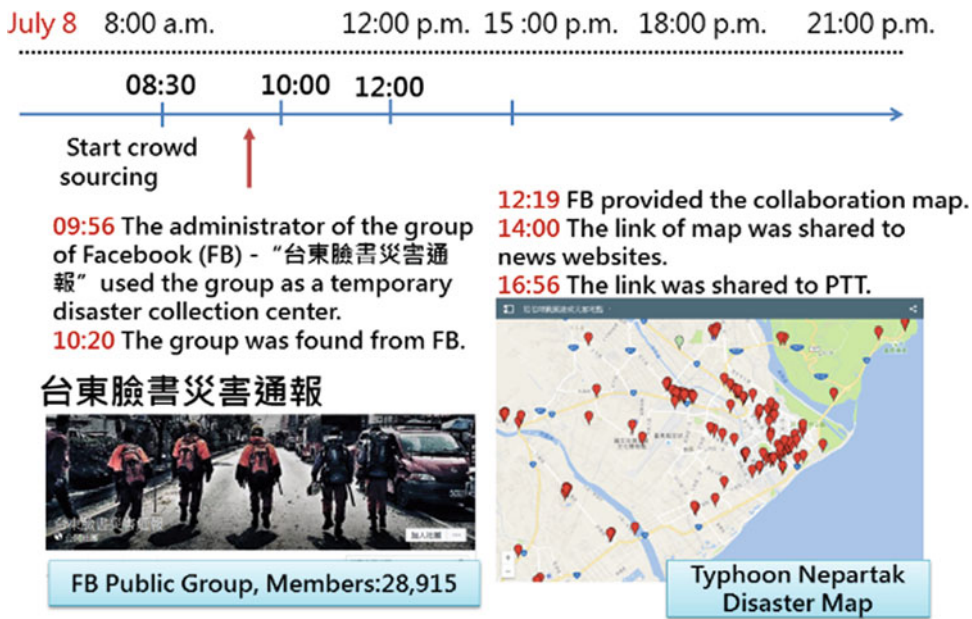
Technology Center for Disaster Reduction (NCDR). It provides users upload the disaster information with web interface during disasters. The user clicks on the map to locate the location, pick up the disaster category, write related descriptions, and upload photos to complete the disaster report (Fig. 6).

The citizen cooperation map uses Facebook-like publishing approach to share disaster information. It is more intuitive to use. In addition to collecting disaster information reported by users, the map can also share the information to Facebook and Google+ for friends or family.

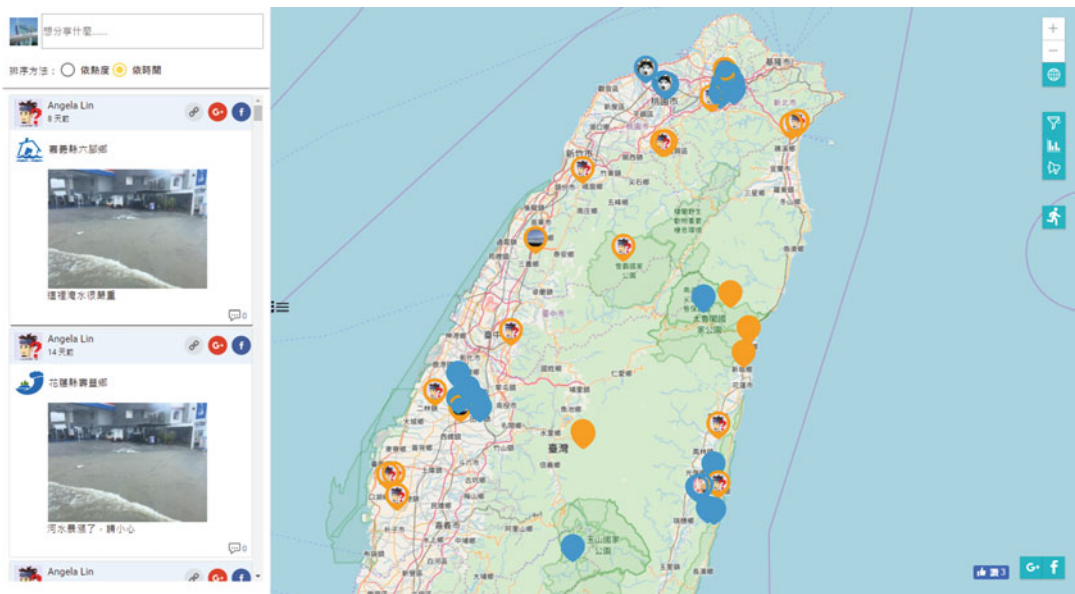
Through citizen cooperation map, disaster prevention personnel of government can extract disaster information into Web-GIS based map. The main purpose of building this map is to provide high-quality information to the commander of center emergency operation central (CEOC) (Fig. 7).

Conclusions

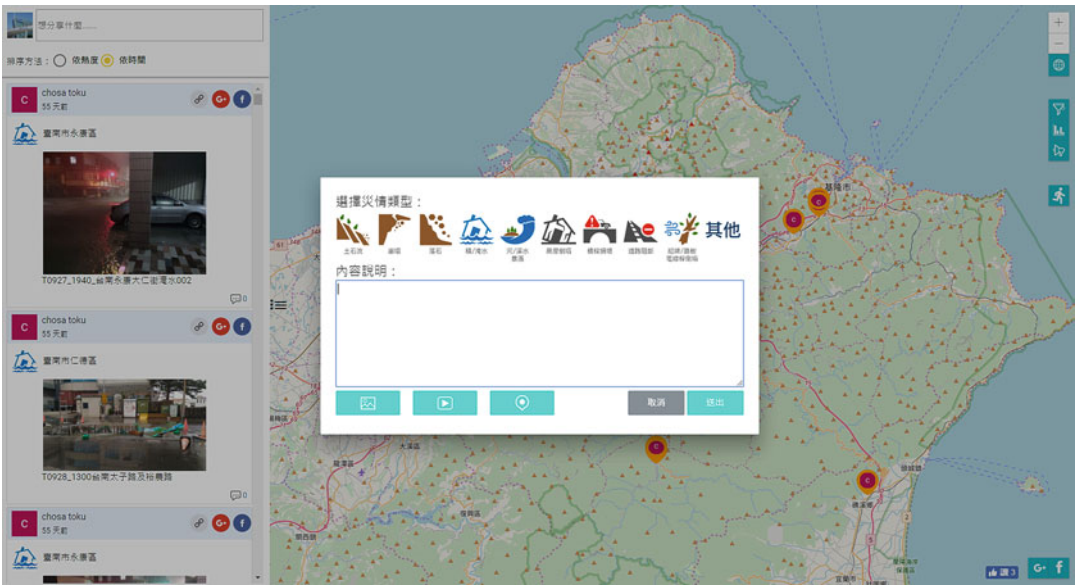
Due to popularity of mobile devices and the development of the Internet, people exchange information through social networking sites has become more frequent. When a disaster occurs, disaster information can be quickly and massively transmitted. The web map is good tools to recruit people to gather disaster information



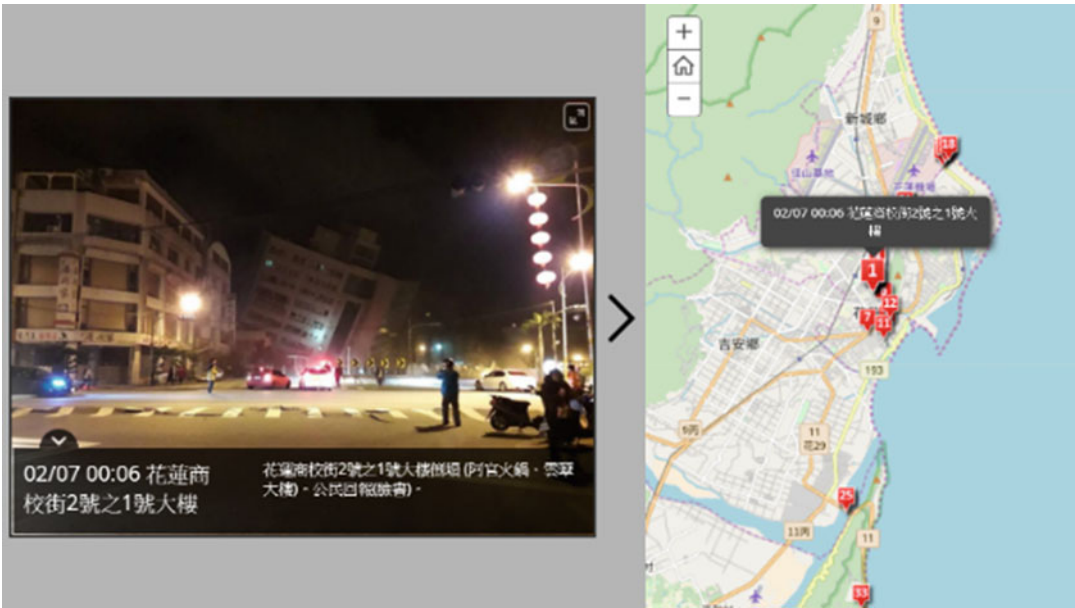
Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 4 Crisis map initiated by local leader



Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 5 The citizen cooperation map



Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 6 Disaster Information Report of citizen cooperation map



Social IoT Crowd-Sourcing on Disaster Reduction, Fig. 7 Web-GIS based map for CEOC

effectively from social network. That disaster information can be immediately gathered for disaster response task to assist the disaster response task.

Cross-References

- Big Data Networking Infrastructure
- Wireless Sensor Network

References

- 2010 Haiti earthquake. Available at https://en.wikipedia.org/wiki/2010_Haiti_earthquake
- 2010 Haiti earthquake damage map. Available at https://wiki.openstreetmap.org/wiki/File:Haiti_earthquake_damage_map.png
- Atzori L, Iera A, Morabito G, Nitti M (2012) The social internet of things (siot)—when social networks meet the internet of things: concept, architecture and network characterization. *Comput Netw* 56(16):3594–3608
- Houston JB, Hawthorne J, Perreault MF, Park EH, Hode MG, Halliwell MR, Turner McGowen SE, Davis R, Vaid S, McElderry JA, Griffith SA (2014) Social media and disasters: a functional framework for social media use in disaster planning, response, and research. 22 September. Available at <https://doi.org/10.1111/disa.12092>
- Karger DR, Oh S, Shah D (2014) Budget-optimal task allocation for reliable crowdsourcing systems. *Oper Res* 62:1–24
- Lina W-Y, Wua T-H, Tsai M-H, Hsua W-C, Choua Y-T, Kanga S-C (2018) Filtering disaster responses using crowdsourcing. *Autom Constr* 91:182–192
- Prpića J, Shuklab PP, Kietzmann JH, McCarthy IP (2015) How to work a crowd: developing crowd capital through crowdsourcing. *Bus Horiz* 58(1):77–85
- Serrat O (2017) Social network analysis. In: *Knowledge solutions*, Springer, Singapore, pp 39–43
- Typhoon Haiyan. Available at https://en.wikipedia.org/wiki/Typhoon_Haiyan
- Typhoon Haiyan/Yolanda Crisis and Relief Map. Available at <https://google.com/crisismap/a/gmail.com/TyphoonYolanda>
- Typhoon Nepartak (2016). Available at [https://en.wikipedia.org/wiki/Typhoon_Nepartak_\(2016\)](https://en.wikipedia.org/wiki/Typhoon_Nepartak_(2016))
- Typhoon Yolanda Map. Available at <https://yolandadata.org/maps/71>
- Xu Z, Liu Y, Yen N, Nei L et al (2016) Crowdsourcing based description of urban emergency events using social media big data. *IEEE Trans Cloud Comput* 99:1
- Zhao Y, Zhu Q (2014) Evaluation on crowdsourcing research: current status and future direction. *Inf Syst Front* 16(3):417–434

Social Network Service and the Internet of Things

- [Social Internet of Things](#)

Social Optimality

- [Efficiency and Pareto Optimality](#)

Soft-Computing

- [Extending WSN Lifetime Based on Evolutionary Clustering Algorithm](#)

Software Platforms for WSN

- [Middleware for Wireless Sensor Networks](#)

Software Router

- [Performance Modeling of Virtual Packet Forwarding Functions for Wireless Network Virtualization](#)

Software-Defined Data Center

- [SDN-Enabled Data Center Networks](#)

Software-Defined Wireless Networking

- [Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks](#)

Space-Shift Keying

- [Spatial Modulation](#)

Space-Time Block Codes

Weifeng Su
Department of Electrical Engineering, State University of New York (SUNY) at Buffalo, Buffalo, NY, USA

Synonyms

[Space-time block coding](#); [Space-time coding and modulation](#)

Definition

Space-time block codes are coding and modulation schemes for multi-input multi-output wireless communication systems which map information bits into multidimensional signals to be transmitted by multiple transmit antennas.

Historical Background

In the 1990s, a revolutionary theoretical work revealed that multi-input multi-output (MIMO) wireless communications can provide substantial capacity increase in rich scattering and reflection environments compared to conventional single-input single-output (SISO) systems (Telatar 1995; Foschini and Gans 1998). Then, performance analysis for MIMO systems with practical signal transmissions and receiver detection/decoding was carried out in Guey et al. (1999), Tarokh et al. (1998) in which two design criteria (one related to coding gain and one related to diversity order or degree of redundancy) were proposed for space-time (ST) signal/code designs in order to minimize the MIMO transceiver error probability. Following the two code design criteria, in the late 1990s and early 2000s, a large number of works on ST signal/code designs were published in literature including orthogonal ST block codes (Alamouti 1998; Tarokh et al. 1999; Ganesan and Stoica 2001; Tirkkonen and Hottinen 2002; Su and Xia 2003; Su et al. 2004), quasi-orthogonal ST block codes (Jafarkhani 2001; Sharma and Papadias 2003; Su and Xia 2002, 2004), cyclic ST block codes (Hochwald and Sweldens 2000; Hughes 2000), unitary ST block codes (Hochwald and Marzetta 2000; Hochwald et al. 2000; Shokrollahi et al. 2001; Hassibi and Hochwald 2002), algebraic ST block codes (Damen et al. 2002; Damen and Beaulieu 2003; Xin et al. 2003), Golden code (Belfiore et al. 2005), perfect ST block codes (Oggier et al. 2006), and so on. Some survey on various STBC code designs and performance comparisons can be found in the books (Larsson and Stoica 2003; Liu et al. 2009) and the references therein. Note that MIMO

technology with STBC has been used in various wireless communication systems including WiFi local area networks and 4G wireless cellular networks.

Foundations

MIMO wireless communication systems with multiple transmit antennas and multiple receive antennas inherently provide a variety of channel links between a transmitter and a receiver in which the received signal at each transmit antenna is a superposition of the transmission signals from all transmit antennas. In typical wireless communication environments, due to signal scattering and reflection, the wireless channels between transmit antennas and receive antennas exhibit random fading with dynamic channel amplitude variation which is often characterized by Rayleigh distribution. The basic idea of coding and modulation for MIMO systems is to encode and map information bits into transmission signals across all transmit antennas. As long as one channel link is strong enough, there is possible to recover the information bits at the receiver side, and the chance that all channel links are weak is relatively low compared to the conventional SISO scenario.

The performance criteria of space-time coding and modulation were thoroughly developed by Guey et al. (1999) and Tarokh et al. (1998). In order to minimize the error probability of the MIMO signal transmission and decoding, space-time codes should be designed following two criteria as follows:

- *Diversity Criterion:* For any two distinct ST code words C and $\tilde{C}$, the rank of the difference matrix $C - \tilde{C}$ should be as large as possible. If the difference matrix between any two distinct code words is of full rank, we say the ST code achieves the full diversity.
- *Coding Gain Criterion:* The minimum value of the product $\prod_{i=1}^{\gamma} \lambda_i$ over all distinct ST code words C and $\tilde{C}$ should be as large as possible, in which λ_i are the non-zero eigenvalues of the difference matrix between two different

code words C and $\tilde{C}$. This quantity is related to the coding gain of the ST code.

The diversity criterion is more important of the two since it determines the slope of the error performance curve. In order to achieve the maximum diversity, the difference matrix $C - \tilde{C}$ has to be full rank for any pair of distinct ST code words C and $\tilde{C}$. The coding gain criterion is of secondary important which should be optimized if the full diversity is achieved. Based on the two code design criteria, a large number of ST signal/code designs have been proposed (see, e.g., Alamouti 1998; Tarokh et al. 1999; Ganesan and Stoica 2001; Tirkkonen and Hottinen 2002; Su and Xia 2002, 2003; Su et al. 2004; Jafarkhani 2001; Sharma and Papadias 2003; Su and Xia 2004; Hochwald and Sweldens 2000; Hughes 2000; Hochwald and Marzetta 2000; Hochwald et al. 2000; Shokrollahi et al. 2001; Hassibi and Hochwald 2002; Damen et al. 2002; Damen and Beaulieu 2003; Xin et al. 2003; Belfiore et al. 2005; Oggier et al. 2006 and the references therein).

An important code design approach is to construct space-time block codes (STBC) from orthogonal designs which have been widely used in practical wireless communication systems including WiFi local area networks and wireless cellular networks. Orthogonal STBC can guarantee to achieve full diversity performance and have simple fast maximum-likelihood (ML) decoding. In 1998, Alamouti proposed the first orthogonal STBC (Alamouti 1998) for MIMO systems with two transmit antennas:

$$G_2(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \\ -x_2^* & x_1^* \end{bmatrix}, \quad (1)$$

in which x_1 and x_2 can be arbitrary complex information symbols with power constraint $E||G_2||_F^2 = 4$. The proposed scheme has an interesting property that for arbitrary complex symbols x_1 and x_2 , the columns of G_2 are orthogonal to each other. Later, Tarokh et al. extended the concept and proposed a series of orthogonal STBC for MIMO systems with

any number of transmit antennas (Tarokh et al. 1999; Ganesan and Stoica 2001; Tirkkonen and Hottinen 2002; Su and Xia 2003; Su et al. 2004). For example, for MIMO systems with $M_t = 3$ or 4 transmit antennas, orthogonal STBC with code rate $R = 3/4$ (i.e., the number of embedded information symbols per time slot) was presented in Tarokh et al. (1999), Ganesan and Stoica (2001), Tirkkonen and Hottinen (2002), and for $M_t = 2^k$ ($k = 1, 2, 3, \dots$); a recursive expression for orthogonal STBC was given in Su and Xia (2003). In general, the orthogonal STBC code designs are related to the theory of orthogonal designs which has a long history in mathematics (Geramita and Seberry 1979). The orthogonality of the resulting STBC code structure ensures that the codes achieve the full diversity order and have the fast ML decoding algorithms.

Note that the code rate of the orthogonal STBC decreases in terms of the number of transmit antennas as $(n_0 + 1)/(2n_0)$ if the number of transmit antennas is $n = 2n_0$ or $n = 2n_0 - 1$ (Su et al. 2004). Due to the code rate limitation of the orthogonal STBC, quasi-orthogonal STBC was proposed in Jafarkhani (2001) to relax the orthogonality requirement in exchange of the rate increase. In the quasi-orthogonal STBC designs, the columns of the code matrix are grouped such that the columns within a group may not be orthogonal, while the columns belonging to different groups are orthogonal to each other. Later, Sharma and Papadias (2003), Su and Xia (2002, 2004), the concept was further extended to design full-diversity quasi-orthogonal STBC in which some of the information symbols are chosen from a rotated version of the conventional symbol constellation with a specifically designed rotation angle. With the moderate decoding complexity increase, the resulting full-diversity quasi-orthogonal STBC can achieve both full code rate and full diversity performance.

A simple yet effective coding scheme to achieve the full space-time diversity is the cyclic STBC which was proposed around 2000 (Hochwald and Sweldens 2000; Hughes 2000):

$$C_l = \sqrt{M_t} \text{diag}\{e^{ju_1\theta_l}, e^{ju_2\theta_l}, \dots, e^{ju_{M_t}\theta_l}\}, \quad (2)$$

where $l = 0, 1, \dots, L - 1$, $\theta_l = \frac{l}{L}2\pi$ and $u_1, u_2, \dots, u_{M_t} \in \{0, 1, \dots, L - 1\}$ are some integers selected to achieve the full diversity and maximize the coding gain. The cyclic STBC has a simple diagonal code structure, and it is a special case of unitary STBC codes. A variety of more general unitary STBC were proposed in literature that can guarantee to achieve the full diversity and have larger coding gain (see, e.g., Hochwald and Marzetta 2000, Hochwald et al. 2000, Shokrollahi et al. 2001, Hassibi and Hochwald 2002 and the reference therein). While the primary application of the unitary STBC is for MIMO differential modulation/demodulation, the resulting codes can also achieve the full diversity and the coding gain of some unitary codes even higher than that of orthogonal STBC. Also, the code structures of the unitary STBC and the design approach were later extended to design more general non-unitary STBC.

Non-unitary STBC code designs have more flexibility in constructing code structures to achieve full diversity and maximize coding gain. An interesting approach is to design algebraic STBC by applying some transforms such as Vandermonde transforms and Hadamard transforms over information symbols to form STBC codes (see, e.g., Damen et al. 2002, Damen and Beaulieu 2003, Xin et al. 2003 and the reference therein). The full diversity of the algebraic STBC is guaranteed based on the algebraic number theory. An enlightening example is the Golden code which was designed based on the Golden number $(1 + \sqrt{5})/2$ for MIMO systems with two transmit antennas (Belfiore et al. 2005). The Golden code achieves the full diversity and has full rate. The concept was also extended to design perfect STBC based on special number fields for MIMO systems with higher number of transmit antennas (Oggier et al. 2006).

Note that most STBC codes were designed for MIMO systems with frequency-nonselective (flat) fading environments. For MIMO broadband wireless communications, the channel may exhibit frequency selective fading, and one may consider space-frequency coding or space-time-frequency coding to exploit the available spa-

tial, temporal, and frequency diversity. Naturally many existing STBC may be applied for broadband MIMO systems; however, the codes have to be modified in certain ways in order to achieve all the full spatial, temporal, and frequency diversity (see, e.g., Bölcskei and Paulraj 2000, Blum et al. 2001, Su et al. 2003, 2005, Liu et al. 2002, Su et al. 2005 and the reference therein).

Key Applications

STBC has been used in various MIMO wireless communication systems including WiFi local area networks, 4G wireless cellular networks, and the emerging 5G wireless networks.

Cross-References

- Massive MIMO
- MIMO Detection
- MIMO Relay
- Network MIMO

References

- Alamouti S (1998) A simple transmit diversity technique for wireless communications. *IEEE JSAC* 16(8): 1451–1458
- Belfiore J-C, Rekaya G, Viterbo E (2005) The golden code: a 2×2 full-rate space-time code with nonvanishing determinants. *IEEE Trans Inf Theory* 51(4): 1432–1436
- Blum R, Li Y, Winters J, Yan Q (2001) Improved space-time coding for MIMO-OFDM wireless communications. *IEEE Trans Commun* 49(11):1873–1878
- Bölcskei H, Paulraj AJ (2000) Space-frequency coded broadband OFDM systems. In: *Proceedings of IEEE WCNC*, Chicago, IL, pp 1–6
- Damen M, Beaulieu NC (2003) On two high-rate algebraic space-time codes. *IEEE Trans Inf Theory* 49(4):1059–1063
- Damen MO, Abed-Meraim K, Belfiore J-C (2002) Diagonal algebraic space-time block codes. *IEEE Trans Inf Theory* 48(3):628–636
- Foschini GJ, Gans MJ (1998) On limits of wireless communications in a fading environment when using multiple antennas. *Wirel Pers Commun* 6:311–335
- Ganesan G, Stoica P (2001) Space-time block codes: a maximum SNR approach. *IEEE Trans Inf Theory* 47:1650–1656

- Geramita AV, Seberry J (1979) Orthogonal designs, quadratic forms and Hadamard matrices. Lecture notes in pure and applied mathematics, vol 43. Marcel Dekker, New York/Basel
- Guey J-C, Fitz MP, Bell MR, Kuo W-Y (1999) Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels. In: Proceedings of IEEE VTC'96, pp 136–140. Also in IEEE Trans Commun 47:527–537
- Hassibi B, Hochwald BM (2002) Cayley differential unitary space-time Codes. IEEE Trans Inf Theory 48(6):1485–1503
- Hochwald BM, Marzetta TL (2000) Unitary space-time modulation for multiple-antenna communication in Rayleigh flat fading. IEEE Trans Inf Theory 46(2):543–564
- Hochwald BM, Sweldens W (2000) Differential unitary space-time modulation. IEEE Trans Commun 48:2041–2052
- Hochwald BM, Marzetta TL, Richardson TJ, Sweldens W, Urbanke R (2000) Systematic design of unitary space-time constellations. IEEE Trans Inf Theory 46(6):1962–1973
- Hughes BL (2000) Differential space-time modulation. IEEE Trans Inf Theory 46:2567–2578
- Jafarkhani H (2001) A quasi-orthogonal space-time block code. IEEE Trans Commun 49(1):1–4
- Larsson EG, Stoica P (2003) Space-time block coding for wireless communications. Cambridge University Press, New York
- Liu Z, Xin Y, Giannakis G (2002) Space-time-frequency coded OFDM over frequency selective fading channels. IEEE Trans Signal Process 50(10):2465–2476
- Liu KJR, Sadek A, Su W, Kwasinski A (2009) Cooperative communications and networking, 1st edn. Cambridge University Press, New York
- Oggier F, Rekaya G, Belfiore J-C, Viterbo E (2006) Perfect space-time block codes. IEEE Trans Inf Theory 52(9):3885–3902
- Sharma N, Papadias CB (2003) Improved quasi-orthogonal codes through constellation rotation. IEEE Trans Commun 51(3):332–335
- Shokrollahi A, Hassibi B, Hochwald BM, Sweldens W (2001) Representation theory for high-rate multiple-antenna code design. IEEE Trans Inf Theory 47:2335–2367
- Su W, Xia X-G (2002) Quasi-orthogonal space-time block codes with full diversity. Proc IEEE GLOBECOM'02 2:1098–1102
- Su W, Xia X-G (2003) On space-time block codes from complex orthogonal designs. Wirel Pers Commun 25(1):1–26. Kluwer Academic Publishers
- Su W, Xia X-G (2004) Signal constellations for quasi-orthogonal space-time block codes with full diversity. IEEE Trans Inf Theory 50(10):2331–2347
- Su W, Safar Z, Olfat M, Liu KJR (2003) Obtaining full-diversity space-frequency codes from space-time codes via mapping. IEEE Trans Signal Process (Special Issue MIMO Wirel Commun) 51(11):2905–2916
- Su W, Xia X-G, Liu KJR (2004) A systematic design of high-rate complex orthogonal space-time block codes. IEEE Commun Lett 8(6):380–382
- Su W, Safar Z, Liu KJR (2005) Full-rate full-diversity space-frequency codes with optimum coding advantage. IEEE Trans Inf Theory 51(1):229–249
- Su W, Safar Z, Liu KJR (2005) Towards maximum achievable diversity in space, time and frequency: performance analysis and code design. IEEE Trans Wirel Commun 4(4):1847–1857
- Tarokh V, Seshadri N, Calderbank AR (1998) Space-time codes for high data rate wireless communication: performance criterion and code construction. IEEE Trans Inf Theory 44(2):744–765
- Tarokh V, Jafarkhani H, Calderbank AR (1999) Space-time block codes from orthogonal designs. IEEE Trans on Inf Theory 45(50):1456–1467
- Telatar IE (1995) Capacity of multi-antenna Gaussian channels. AT&T Bell Labs, Technical Report
- Tirkkonen O, Hottinen A (2002) Square-matrix embeddable space-time block codes for complex signal constellations. IEEE Trans Inf Theory 48(2):384–395
- Xin Y, Wang Z, Giannakis G (2003) Space-time diversity systems based on linear constellation precoding. IEEE Trans Wirel Commun 2(2):294–309

Space-Time Block Coding

► Space-Time Block Codes

Space-Time Coding and Modulation

► Space-Time Block Codes

Space-Time Shift Keying

► Spatial Modulation

Sparse Code Multiple Access

Ziyang Li and Wen Chen
Shanghai Institute of Advanced
Communications and Data Sciences, Department
of Electronic Engineering, Shanghai Jiao Tong
University, Shanghai, China

Synonyms

Sparse spreading multiple access

Definition

Sparse code multiple access (SCMA) is a nonorthogonal multiple access technique based on codebooks of sparse structure.

Historical Background

Future 5G wireless networks are expected to support massive connectivity, higher throughput, lower latency, and lower controlling signal overhead. Compared with the conventional orthogonal multiple access techniques such as *code division multiple access (CDMA)* and *orthogonal frequency division multiple access (OFDMA)* utilized in current networks, *nonorthogonal multiple access (NOMA)*, which has the advantage of higher resource utilization, is more promising in 5G network.

Motivated by the design of CDMA chip sequences, the idea of sparse spreading, called *Low Density Signature (LDS)*, was first proposed by Reza Hoshyar (Hoshyar et al. 2008). LDS is a special case of multicarrier CDMA with a few nonzero elements in a long spreading signature. This sparsity characteristic can reduce the complexity of the multiuser detection called *Message Passing Algorithm (MPA)* (Razavi et al. 2011, 2012). In an LDS combined with Orthogonal Frequency Division Multiplexing (OFDM) system, which is called LDS-OFDM

(Hoshyar et al. 2010), the capacity is proved to have a superior performance than OFDMA.

To meet the requirements of 5G, *Sparse Code Multiple Access (SCMA)* (Nikopour and Baligh 2013), a new nonorthogonal codebook-based multiple access method, is proposed by Huawei in 2013. SCMA can be seen as a generalization of LDS. In LDS, incoming bits are mapped to a QAM symbol, and the repetitions of the QAM symbol are transmitted through the subcarriers according to the designed signature. But in SCMA, incoming bits are directly mapped to multidimensional complex codewords selected from predefined codebook (Taherzadeh et al. 2014). Instead of simple repetition of QAM symbol in LDS, SCMA provides significant shaping gain with the multidimensional constellation design (Boutros et al. 1996). Similar to LDS, only a small number of dimensions are used to transmit data; therefore, the codewords of SCMA are sparse. In the SCMA codebook, the nonzero dimensions are corresponding to the subcarriers in use. Besides, SCMA can also utilize iterative MPA detection with near optimal performance in the MAP sense.

The current research studies of SCMA are mainly focused on *capacity analysis*, *codebook design*, *low-complexity decoding*, *resource allocation*, *blind detection*, and so on. In *capacity analysis* area, researchers compare the capacity of SCMA system with the LDS system (Cheng et al. 2015) and analyze the capacity of downlink massive MIMO MU-SCMA system (Liu et al. 2015). In *codebook design* area, the method of *Shuffling* (Taherzadeh et al. 2014) and the method based on star-QAM (Yu et al. 2015) are proposed to decrease the *Bit Error Rate (BER)*. In *low-complexity decoding* area, some algorithms like PM (Mu et al. 2015), Quasi-ML (Zou et al. 2016), List Sphere Decoding (Wei and Chen 2017) are proposed to reduce the decoding complexity. In *resource allocation* area, the SCMA downlink power allocation to maximize weighted sum-rate has been studied in Nikopour et al. (2014) and uplink resource allocation based on a special structure codebook studied in Li et al. (2016). In *blind detection* area, contention-based blind

detection is proposed to support Grant-free transmission (Bayesteh et al. 2014; Au et al. 2014).

Foundations

SCMA system mainly consists of two parts: SCMA encoder and SCMA decoder.

SCMA encoder can be defined as a mapping from $\log_2(M)$ bits to a K -dimensional codeword of size M selected from a predefined codebook. K dimensions are corresponding to K different orthogonal tones, such as OFDMA subcarriers. The K -dimensional codeword is a vector with only $N < K$ nonzero entries. Users cannot transmit data through the subcarriers represented by the other $N - K$ zero entries. Theoretically, each user can be allocated to more than one codebook, and each codebook can be utilized by more than one user generally.

Figure 1 shows an example of an SCMA encoder, with six layered codebooks (*variable nodes*) and four subcarriers (*function nodes*). Each row denotes a dimension, and each column means a 4-dimensional codeword. In each codebook, the constellation size is 4, which means there are four different codewords can be chosen. The white entries denote the zero elements and the colored entries denote the nonzero elements in the codebooks. For example, in Codebook 1, the entries in the first row is colored and the entries in the third row is white, which means the first dimension is nonzero and third dimension is zero. In each codebook, there are two nonzero dimensions with colored lattice. In an AWGN channel, the signal received in the base station is the superposition of the codewords selected from the codebooks.

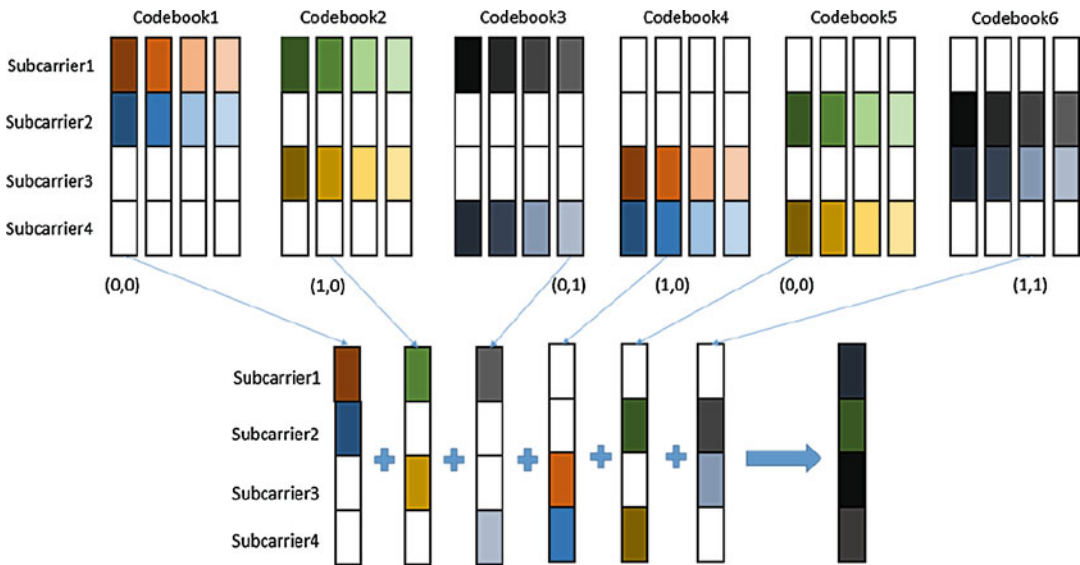
Codebook design is the most important part in SCMA encoder. The target is to design a multidimensional lattice constellation with dimensional dependency and power variation of the constellation while maintaining large minimum Euclidean distance. Generally, there are three stages to design SCMA code (Nikopour and Baligh 2013; Taherzadeh et al. 2014): (1) *Mapping Matrix* determines the number of layers interfering at each subcarrier, which represents

the complexity of MPA detection. For example, Fig. 1 can be considered as a *mapping matrix*, which means that each layer will be interfered by two other layers. (2) *Constellation Points and Multidimensional Mother Constellation* design. First, design a base constellation with a maximized minimum Euclidean distance. Second, a unitary rotation, which might be designed to maximize the minimum product distance of the constellation, can be applied on the base constellation to control the dimensional dependency and power variation. Third, build the complex constellation using the rotated base constellation by shuffling. Last, utilize the rotation to minimize the projection points. (3) *Constellation Function Operator*, which includes several operators like *complex conjugate*, *phase rotation*, and *dimensional permutation*, aims to design distinct codebooks for the collision layers.

Currently, SCMA decoder usually utilizes MPA detection or improved version of MPA detection. There are three steps of the conventional MPA detection: (1) *Initialize the conditional probability*. For the SCMA system in Fig. 1, in each resource node, there are three layers collided and each layer has four possibilities of constellation points. Then there are $4 * 4 * 4$ combinations of transmitted signals which determine the number of conditional probabilities. (2) *Iterative message passing through the variable nodes and function nodes*. Function node passes the updated message to its neighboring variable nodes, then the variable node passes the updated message to its neighboring function node, which constitutes an iteration. (3) *Log-Likelihood-Rate(LLR) calculation*. After several iterations, LLR output can be calculated by the probability guess of codeword at each layer.

Key Applications

Sparse code multiple access technique will be utilized in future 5G networks, which will be applied mainly in three scenarios: (1) *Enhanced mobile broadband (eMBB)*, for example, Ultra HD video transmission



Sparse Code Multiple Access, Fig. 1 SCMA encoder

and High Speed Mobile Broadband Wireless Communication. (2) *Ultra-reliable and low latency communications (uRLLC)*, for example, unmanned and automatic driving. (3) *massive machine type of communication (mMTC)*, for example, Internet of things (IoT).

Cross-References

- [5G Wireless](#)
- [Internet of Things](#)
- [Multiple Access in High Frequency](#)
- [Non-orthogonal Multiple Access \(NOMA\)](#)
- [Sparse Code Multiple Access](#)
- [Vehicle-to-Vehicle Communications](#)

References

- Au K, Zhang L, Nikopour H, Yi E, Bayesteh A, Vilaipornsawai U, Ma J, Zhu P (2014) Uplink contention based scma for 5g radio access. In: Globecom workshops (GC Wkshps), 2014. IEEE, Austin, US pp 900–905
- Bayesteh A, Yi E, Nikopour H, Baligh H (2014) Blind detection of SCMA for uplink grant-free multiple access. In: International symposium on wireless communications systems. IEEE, Barcelona, Spain pp 853–857
- Boutros J, Viterbo E, Rastello C, Belfiore J-C (1996) Good lattice constellations for both Rayleigh fading and Gaussian channels. *IEEE Trans Inf Theory* 42(2): 502–518
- Cheng M, Wu Y, Chen Y (2015) Capacity analysis for non-orthogonal overloading transmissions under constellation constraints. In: International conference on wireless communications & signal processing. IEEE, Nanjing China pp 1–5
- Hoshyar R, Wathan FP, Tafazolli R (2008) Novel low-density signature for synchronous CDMA systems over AWGN channel. *IEEE Trans Signal Process* 56(4):1616–1626
- Hoshyar R, Razavi R, Al-Imari M (2010) LDS-OFDM an efficient multiple access technique. In: Proceedings of the IEEE 71st vehicular technology conference (VTC 2010-Spring), 2010. IEEE, Taipei, Taiwan pp 1–5
- Li Z, Chen W, Wei F, Wang F, Xu X, Chen Y (2016) Joint codebook assignment and power allocation for SCMA based on capacity with Gaussian input. In: IEEE/CIC international conference on communications in China. IEEE, Chengdu, China pp 1–6
- Liu T, Li X, Qiu L (2015) Capacity for downlink massive mimo muscma system. In: International conference on wireless communications & signal processing. IEEE, Nanjing China pp 1–5
- Mu H, Ma Z, Alhaji M, Fan P, Chen D (2015) A fixed low complexity message pass algorithm detector for up-link

SCMA system. *IEEE Wireless Commun Lett* 4(6):585–588

Nikopour H, Baligh H (2013) Sparse code multiple access. In: 2013 IEEE 24th international symposium on personal indoor and mobile radio communications (PIMRC). IEEE, London UK pp 332–336

Nikopour H, Yi E, Bayesteh A, Au K, Hawryluck M, Baligh H, Ma J (2014) SCMA for downlink multiple access of 5g wireless networks. In: Global communications conference (GLOBECOM), 2014 IEEE. IEEE, Austin USA pp 3940–3945

Razavi R, Imran MA, Tafazolli R (2011) Exit chart analysis for turbo LDS-OFDM receivers. In: Proceedings of the 7th international wireless communications and mobile computing conference (IWCMC), 2011. IEEE, Istanbul, Turkey pp 354–358

Razavi R, Al-Imari M, Imran MA, Hoshyar R, Chen D (2012) On receiver design for uplink low density signature OFDM (LDS-OFDM). *IEEE Trans Commun* 60(11):3499–3508

Taherzadeh M, Nikopour H, Bayesteh A, Baligh H (2014) SCMA codebook design. In: 2014 IEEE 80th vehicular technology conference (VTC Fall). IEEE, Vancouver, Canada pp 1–5

Wei F, Chen W (2017) Low complexity iterative receiver design for sparse code multiple access. *IEEE Trans Commun* 65(2):621–634

Yu L, Lei X, Fan P, Chen D (2015) An optimized design of SCMA codebook based on star-QAM signaling constellations. In: International conference on wireless communications & signal processing. IEEE, Nanjing China pp 1–5

Zou J, Zhao H, Li X (2016) A low-complexity tree search based quasi-ml receiver for SCMA system. In: IEEE international conference on computer and communications. IEEE, Chengdu China pp 319–323

Spatial Modulation

Shinya Sugiura

Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Koganei, Japan

Synonyms

[Index modulation](#); [Permutation modulation](#); [Space-shift keying](#); [Space-time shift keying](#)

Definitions

Spatial modulation is a single-radio-frequency (RF) multiple-antenna-assisted information transmission technique that combines two types of modulation schemes: the classic amplitude and phase shift keying (APSK), and antenna index modulation. More specifically, at a transmitter, one of multiple transmit antenna elements is selected depending on a subset of the input bits, and then an APSK symbol modulated based on the remaining input bits is transmitted from the single selected transmit antenna element. Hence, both the indices of the activated antenna element as well as the APSK-modulated symbol convey information to the receiver. Since only a single antenna element is activated during each symbol interval, spatial modulation allows us to maintain a cost/energy-efficient single-RF transmitter structure despite the use of multiple transmit antenna elements.

Sparse Signal Processing

► [Big Data in 5G](#)

Sparse Spreading Multiple Access

► [Sparse Code Multiple Access](#)

Historical Background

Spatial modulation is one of the multiple-input multiple-output (MIMO) transmission techniques (Di Renzo et al. 2011, 2014, 2016; Sugiura et al. 2012; Yang et al. 2016), whereby multiple antennas are equipped at a transmitter and/or a receiver, used to enhance the achievable performance. Depending on the target gain attainable by multiple antennas, the

classic MIMO schemes are classified into four categories, namely, spatial multiplexing, spatial diversity, beamforming, and space division multiple access (SDMA), which allow us to achieve a rate enhancement, an increased reliability, a high signal-to-noise ratio, and a large number of supported users, respectively (Hanzo et al. 2009; Sugiura et al. 2012). In contrast with these four classic MIMO arrangements, spatial modulation has been recently recognized as a fifth MIMO concept because it achieves exclusive benefits and provides further design degrees of freedom, owing to its specific sparse use of transmit antenna elements.

In 2001 (Chau and Yu 2001), the original concept of spatial modulation was proposed, whereby the combination of activated transmit antenna elements per symbol interval conveys information bits. However, multiple RF chains are needed in this original arrangement of spatial modulation, and hence no benefits were attainable over the conventional MIMO techniques. However, in Haas et al. (2002), the modulation principle of Chau and Yu (2001) was modified to one relying on the activation of only a single transmit antenna element in each symbol interval, which enables a single-RF transmitter structure. This has become the basic principle of spatial modulation.

Since then, diverse spatial modulation related techniques have been proposed. For example, spatial modulation was combined with the other four MIMOs in order to simultaneously amalgamate multiple of the five MIMOs' functions. In Jeganathan et al. (2008), spatial multiplexing and spatial modulation were combined for the sake of attaining a high transmission rate while maintaining a reduced number of RF chains. In Sugiura et al. (2010) and Di Renzo and Haas (2013), spatial diversity and spatial modulation were used together for the sake of attaining both reliability and rate enhancement through a single RF transmitter. Additionally, the issues of channel estimation (Sugiura and Hanzo 2012), non-orthogonal multiple access (Wang et al. 2017), differential modulation (Bian et al. 2013), reconfigurable-antenna-based modulation (Sugiura 2014; Boudida et al. 2016), and cooperative

communications (Sugiura et al. 2011a) have been investigated in the context of spatial modulation.

Foundations

Transmitter Model

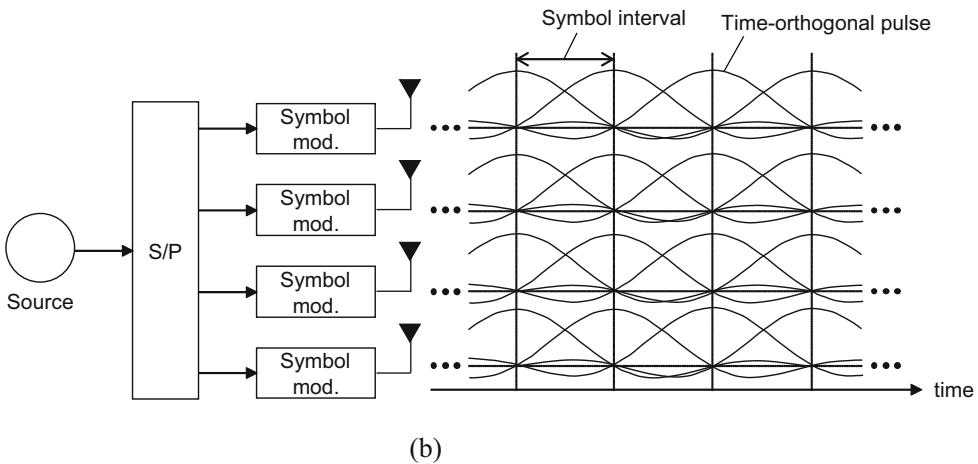
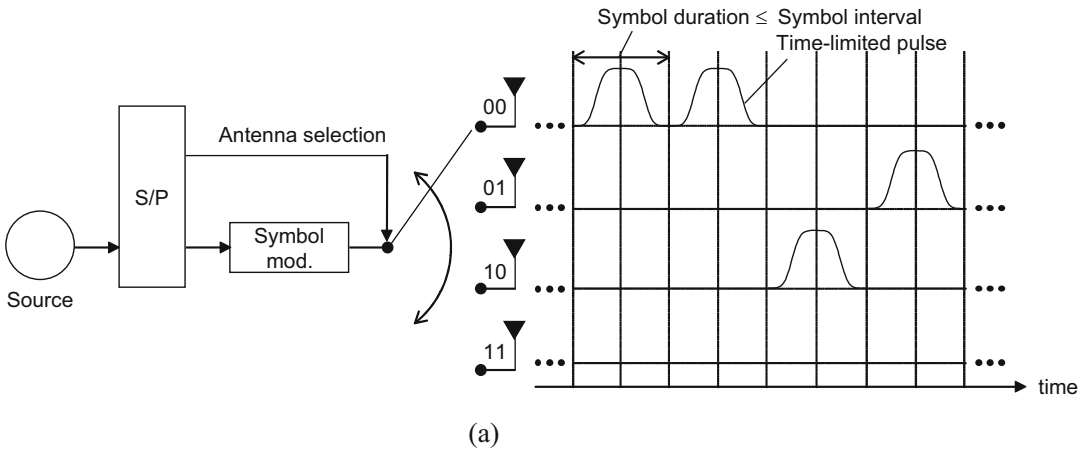
Figure 1a portrays the transmitter model of spatial modulation. We assume M transmit antenna elements are equipped and L -size APSK constellations. At the spatial modulation transmitter, a single APSK-modulated symbol is transmitted during each symbol interval while typically assuming the use of a single-carrier transmission. Note that in order to maintain a single-RF transmitter structure, the symbol duration has to be less than the symbol interval (Sugiura et al. 2017). Importantly, spatial modulation is characterized by the sparse use of transmit antenna elements, while the four classic full-RF MIMO schemes rely on the simultaneous use of all/multiple transmit antenna elements, as shown in Fig. 1b.

More specifically, in each symbol interval, $B = \log_2 M + \log_2 L$ information bits are input to the spatial modulation encoder. Here, one of M indexed transmit antenna elements is selected based on the $\log_2 M$ input bits, and the $\log_2 L$ input bits are modulated onto an APSK symbol. Finally, the APSK-modulated symbol is transmitted from the activated transmit antenna element. Hence, the normalized transmission rate of spatial modulation is formulated by

$$B = \log_2 M + \log_2 L \quad [\text{bits/symbol}], \quad (1)$$

which increases logarithmically with an increase in the number of transmit antenna elements. The transmission rate (1) of spatial modulation is typically lower than that of spatial multiplexing, i.e., $M \log_2 L$ bits/symbol, when considering the same number of transmit antenna elements. This is because the transmission rate of full-RF spatial multiplexing increases linearly with the number of transmit antenna elements.

Since only a single transmit antenna is used in each symbol interval, the spatial modulation



Spatial Modulation, Fig. 1 Transmitter structures: (a) single-RF spatial modulation and (b) full-RF spatial multiplexing. © 2017 IEEE. (Reprinted, with permission, from Sugiura et al. 2017)

transmitter needs only a single RF chain, which is a specific benefit of spatial modulation. Note that, as shown in Fig. 1b, a classic MIMO transmitter typically imposes the requirement of a full-RF structure, where the number of RF chains is the same as that of transmit antenna elements. As mentioned in Di Renzo et al. (2014), the dynamic RF power consumption, the static power consumption imposed by the circuit (e.g., the power amplifier), increases linearly with the number of transmit antenna elements in a conventional full-RF MIMO transmitter. Naturally, the transmitter cost can easily be made low for a low number of RF chains. Hence, spatial modulation is espe-

cially beneficial for the massive MIMO scenario, where a substantially large number of transmit antenna elements are used.

Receiver Model

Consider the receiver to be equipped with N receive antenna elements. By assuming a single-carrier transmission over a frequency-flat fading channel for the sake of simplicity, the received signal model of the spatial modulation system is represented by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \quad (2)$$

where $\mathbf{y} \in \mathcal{C}^N$ are the received signals, and $\mathbf{H} \in \mathcal{C}^{N \times M}$ are the MIMO channel coefficients. Furthermore, $\mathbf{s} \in \mathcal{C}^M$ is the transmit symbol vector, where there is only a single non-zero element, corresponding to an activated transmit antenna element. Also, $\mathbf{n} \in \mathcal{C}^N$ comprises the associated additive white Gaussian noise (AWGN) components, which are represented by random variables obeying the zero-mean complex-valued Gaussian distribution having variance of N_0 .

Note that, while in spatial multiplexing the effects of inter-channel interference (ICI) are imposed at the receiver, the spatial modulation receiver does not suffer from the effects of ICI, as an explicit benefit of the spatial modulation scheme's principle of a single transmit antenna activation. This allows us to benefit from low-complexity detection algorithms (Sugiura et al. 2011b), which are suitable for operation in mobile handsets.

While in most of the detection algorithms developed for spatial modulation frequency-flat fading scenarios were assumed, some practical equalizers of spatial modulation were proposed for frequency-selective fading scenarios (Rajashekar et al. 2013; Sugiura and Hanzo 2015).

Fundamental Limitations

Against the above-mentioned benefits specific to spatial modulation, there are some fundamental limitations imposed (Sugiura et al. 2017), which remain open issues.

- *Pulse shaping:* A spatial modulation transmitter typically has to transmit a whole symbol during a symbol duration, due to the symbol-wise antenna switching principle, as shown in Fig. 1a. Hence, time-orthogonal bandwidth-efficient shaping filters, such as root raised cosine filters, are not available under spatial modulation. However, most of the previous studies on spatial modulation do not take into account this limitation. As an exceptional example, in order to reduce the effects of this information-theoretic penalty imposed by time-limited pulse shaping, the use of two RF

chains, rather than a single one, was proposed in Ishibashi and Sugiura (2014).

- *Single-carrier transmission:* Spatial modulation does not have any compelling gain over the conventional single-antenna-based system when used with orthogonal frequency-division multiplexing (OFDM) (Sugiura and Hanzo 2015). This is because, in a single-RF spatial modulation transmitter, subcarrier-wise antenna selections are not practical, and hence only frame-wise antenna selection is possible. However, OFDM typically employs at least hundreds of subcarriers, and the rate enhancement attained by spatial modulation becomes negligible in such a scenario. For this reason, spatial modulation clearly has to employ single-carrier transmissions rather than multi-carrier ones. Furthermore, since the current wireless standards assume broadband communications, single-carrier transmissions typically suffer from frequency-selective channels, experiencing a long-tap delay, hence requiring the use of efficient equalizers, such as those of Rajashekar et al. (2013), Sugiura and Hanzo (2015), at the receiver.

Key Applications

As mentioned above, spatial modulation is characterized by a high energy efficiency and a low cost, owing to a single-RF MIMO transmitter structure, as well as an ICI-free low-complexity receiver structure. Therefore, key applications of spatial modulation are expected in both uplink and downlink scenarios in cellular systems, as follows:

- *Downlink scenario (Sugiura et al. 2017):* Since a base station has the capability of having a large-scale antenna array installed in a next-generation of wireless systems, spatial modulation is beneficial for high-rate spatial-modulation-based symbol transmission. Furthermore in general, spatial modulation operates in an open-loop manner, whereby a transmitter does not need any feedback associated with channel state information from the receiver. Hence, frequency-division

duplex as well as time-division duplex can be used for massive-MIMO spatial-modulation downlink, unlike a massive MIMO system based on spatial multiplexing (Larsson et al. 2014). More importantly, in the massive MIMO based on spatial modulation, the static energy consumed by an RF circuit is kept low, owing to the single-RF transmitter structure. Furthermore, since signals received at a receiver are free from ICI, a low-complexity detection algorithm suitable for a mobile handset becomes possible.

- *Uplink scenario* (Wang et al. 2015): Spatial modulation is also beneficial for an uplink scenario where multiple mobile terminals communicate with a base station equipped with a large number of antenna elements. Since each mobile terminal needs only a single RF chain in spatial modulation, despite the use of multiple transmit antennas, spatial modulation is useful for maintaining a simplified mobile handset architecture, unlike the conventional full-RF MIMO one.

Moreover, when considering the use of high carrier frequencies, such as in the millimeter wave case, the antenna size and separation both become significantly small, and hence the installation of a large-scale antenna array becomes realistic. This is especially beneficial for spatial modulation, since an exponential increase in the number of transmit antenna elements is needed for a linear increase in the transmission rate.

Furthermore, the use of spatial modulation in visible light communications has been also considered (Fath and Haas 2013). This is for the sake of increasing the spectral efficiency, because in general, the modulation bandwidth of light-emitting diodes is limited.

Cross-References

- [Index Modulation](#)
- [Massive MIMO](#)
- [MIMO Detection](#)

References

- Bian Y, Wen M, Cheng X, Poor HV, Jiao B (2013) A differential scheme for spatial modulation. In: IEEE global communications conference, Atlanta, pp 3925–3930
- Bouida Z, El-Sallabi H, Ghayeb A, Qaraqa KA (2016) Reconfigurable antenna-based space-shift keying (SSK) for MIMO Rician channels. *IEEE Trans Wireless Commun* 15(1):446–457
- Chau YA, Yu SH (2001) Space modulation on wireless fading channels. In: IEEE 54th vehicular technology conference, Atlantic City, vol 3, pp 1668–1671
- Di Renzo M, Haas H (2013) On transmit diversity for spatial modulation MIMO: impact of spatial constellation diagram and shaping filters at the transmitter. *IEEE Trans Veh Technol* 62(6):2507–2531
- Di Renzo M, Haas H, Grant PM (2011) Spatial modulation for multiple-antenna wireless systems: a survey. *IEEE Commun Mag* 49(12):182–191
- Di Renzo M, Haas H, Ghayeb A, Sugiura S, Hanzo L (2014) Spatial modulation for generalized MIMO: challenges, opportunities and implementation. *Proc IEEE* 102(1):1–47
- Di Renzo M, Haas H, Ghayeb A, Hanzo L, Sugiura S (2016) Spatial modulation for multiple-antenna communication. In: Wiley encyclopedia of electrical and electronics engineering. Wiley, New York
- Fath T, Haas H (2013) Performance comparison of MIMO techniques for optical wireless communications in indoor environments. *IEEE Trans Commun* 61(2):733–742
- Haas H, Costa E, Schulz E (2002) Increasing spectral efficiency by data multiplexing using antenna arrays. In: IEEE international symposium on personal, indoor and mobile radio communications, Lisbon, vol 2, pp 610–613
- Hanzo L, Alamri O, El-Hajjar M, Wu N (2009) Near-capacity multi-functional MIMO systems: sphere-packing, iterative detection and cooperation. Wiley/IEEE Press, Chichester
- Ishibashi K, Sugiura S (2014) Effects of antenna switching on band-limited spatial modulation. *IEEE Wireless Commun Lett* 3(4):345–348
- Jeganathan J, Ghayeb A, Szczecinski L (2008) Generalized space shift keying modulation for MIMO channels. In: IEEE 19th international symposium on personal, indoor and mobile radio communications, Cannes, pp 1–5
- Larsson EG, Edfors O, Tufvesson F, Marzetta TL (2014) Massive MIMO for next generation wireless systems. *IEEE Commun Mag* 52(2):186–195
- Rajashekar R, Hari KVS, Hanzo L (2013) Spatial modulation aided zero-padded single carrier transmission for dispersive channels. *IEEE Trans Commun* 61(6):2318–2329
- Sugiura S, Hanzo L (2012) Effects of channel estimation on spatial modulation. *IEEE Signal Process Lett* 19(12):805–808

- Sugiura S, Hanzo L (2015) Single-RF spatial modulation requires single-carrier transmission: frequency-domain turbo equalization for dispersive channels. *IEEE Trans Veh Technol* 64(10):4870–4875
- Sugiura S, Chen S, Hanzo L (2010) Coherent and differential space-time shift keying: a dispersion matrix approach. *IEEE Trans Commun* 58(11):3219–3230
- Sugiura S, Chen S, Haas H, Grant PM, Hanzo L (2011a) Coherent versus non-coherent decode-and-forward relaying aided cooperative space-time shift keying. *IEEE Trans Commun* 59(6):1707–1719
- Sugiura S, Xu C, Ng SX, Hanzo L (2011b) Reduced-complexity coherent versus non-coherent QAM-aided space-time shift keying. *IEEE Trans Commun* 59(11):3090–3101
- Sugiura S, Chen S, Hanzo L (2012) A universal space-time architecture for multiple-antenna aided systems. *IEEE Commun Surv Tutor* 14(2):401–420
- Sugiura S (2014) Coherent versus non-coherent reconfigurable antenna aided virtual MIMO systems. *IEEE Signal Process Lett* 21(4):390–394
- Sugiura S, Ishihara T, Nakao M (2017) State-of-the-art design of index modulation in the space, time, and frequency domains: benefits and fundamental limitations. *IEEE Access* 5:21774–21790
- Wang S, Li Y, Zhao M, Wang J (2015) Energy-efficient and low-complexity uplink transceiver for massive spatial modulation MIMO. *IEEE Trans Veh Technol* 64(10):4617–4632
- Wang X, Wang J, He L, Song J (2017) Spectral efficiency analysis for downlink NOMA aided spatial modulation with finite alphabet inputs. *IEEE Trans Veh Technol* 66(11):10,562–10,566
- Yang P, Xiao Y, Guan YL, Hari KVS, Chockalingam A, Sugiura S, Haas H, Di Renzo M, Masouros C, Liu Z et al (2016) Single-carrier SM-MIMO: a promising design for broadband large-scale antenna systems. *IEEE Commun Surv Tutor* 18(3):1687–1716

Spatial Uncorrelation

- [Wireless Key Establishment](#)

Spectral Efficiency

- [Transmission Capacity](#)

Spectrum Access

- [Spectrum Resource Allocation in Cognitive Radio Networks](#)

Spectrum Database

- [Incentive Mechanism for Crowdsourcing-Based Spectrum Measurement](#)

Spectrum Hole

- [Spectrum Sensing in Cognitive Radio Networks](#)

Spectrum Resource Allocation

- [Spectrum Resource Allocation in Cognitive Radio Networks](#)

Spectrum Resource Allocation in Cognitive Radio Networks

Xin Liu

Dalian University of Technology, Dalian, China

Synonyms

[Cognitive radio](#); [Spectrum access](#); [Spectrum resource allocation](#)

Definition

In recent years, the spectrum allocation mode based on static authorization has been unable to ever-increasing spectrum demand. There are a lot of unused spectrum resources in both the time domain and frequency domain on a dedicated frequency band. However, some frequency bands are too crowded, such as the mobile phone communication band. It results in the uneven distribution and waste of the spectrum resources (Cabric et al. 2004). The key idea of cognitive radio is to perceive the environment using devices that have communication sensing capabilities. It can perceive and respond to the current communication

environment such as spectrum bandwidth, idle time slot, main user mode, and so on. Moreover, it can make full use of the available spectrum resources to solve the problem of low frequency utilization rate under the traditional distribution mode.

Historical Background

The term cognitive radio was first proposed by Joseph Mitola based on the concept of software radio. In 1999, Mitola described a cognitive radio system in his doctoral thesis, which enhanced the flexibility of personal wireless services through a wireless knowledge description language, extended the cognitive radio, and gave an interesting conceptual summary of the interdisciplinary cognitive radio. He believes that software radio is the ideal platform for implementing cognitive radio (Mitola 2001). In his dissertation, Joseph Mitola had studied the learning and reasoning capabilities of realizing cognitive radio through software at high-level, but there was still a lack of research on the physical layer and link layer of the special wireless architecture. In 2001, Simon Haykin further elaborated cognitive radio from the perspective of signal processing in an invitation. He thought that by changing the operating parameters (such as transmission power, carrier frequency, and modulation technology) of cognitive radio in real time to adapt to the statistical changes of the received radio signals, the two most important goals of cognitive radio: highly reliable communications at any time and any place and efficient utilization of spectrum resources can be achieved. In 2003, the FCC held a cognitive radio seminar to discuss the related technical issues of using cognitive radio technology to achieve flexible spectrum utilization. The FCC calls any radio system with adaptive spectrum sensing capability as cognitive radio and further describes cognitive radio applications and key technologies. At the same time as the theoretical research and development of cognitive radio technology, some scientific research projects related to it have also been

carried out in the world. Some representative examples are:

1. The cognitive radio spectrum pool system proposed by Prof. F. K. Jondral, Karlsruhe University, Germany, etc., is an OFDM-based centralized control dynamic spectrum access system. The system architecture includes base stations and mobile users. The research scenarios are mainly focused on the dynamic sharing of spectrum resources between OFDM wireless LANs (such as IEEE 802.11a/g) and GSM networks. With OFDM, specific sub-carriers can be zeroed, so that interference characteristics can be avoided and the available spectrum resources can be used dynamically. The spectrum pool system implements the perception of network available spectrum resources through a specially designed frame structure.
2. The European DRIVE/Over DRIVE project enables cognitive radio dynamic spectrum sharing among heterogeneous networks by establishing a common coordination channel. At present, the project studies two dynamic spectrum allocation algorithms: dynamic spectrum allocation in time and space domain. Time dynamic spectrum allocation refers to that a radio access network can use spectrum resources not currently used by other radio access networks in specific time; and spatial dynamic spectrum allocation allows spectrum allocation for adapting to service changes in different areas.
3. The American CORVUS system utilizes cognitive radio methods to use virtual unlicensed spectrum. The system was proposed by Professor R. W. Brodersen from the University of California, Berkeley, with the goal of detecting and utilizing spectrum in a coordinated manner. In the CORVUS system, the concept of user grouping is proposed. The utilization of dynamic spectrum in group users can be coordinated by the group control channel. And then, the dynamic spectrum distribution between groups can be coordinated by the common control channel. Finally, the reliable link maintenance protocol under dynamic

spectrum access was proposed. The system is developing test beds to evaluate the performance of physical and media access control layers.

4. The DIMSUM Net system is a network architecture proposed by researchers at Lucent Bell Labs and Stevens Institute of Technology for implementing Spectrum Statistical Multiple Access (SMA) through the Coordinated Access Band (CAB). The system uses the coordinated access frequency band to improve the spectrum access efficiency and fairness, and improves the spectrum utilization through statistical multiple access.

Foundations

In cognitive radios, the MAC layer is responsible for handling spectrum management units, which include spectrum sensing, decision making, sharing, and handover. This section provides an overview of spectrum sharing and allocation methods according to different classification criteria.

1. Classification methods: static allocation, dynamic allocation, and hybrid allocation (Hossain et al. 2009). Static allocation means that the frequency band is allocated to designated users according to authorization, and other unlicensed users cannot use the allocated spectrum resources. This allocation method leads to low spectrum utilization. The advantage is that there is no communication interference between users and the communication of authorized users can be fully guaranteed quality. In the dynamic allocation mode, the system allocates spectrum resources adaptively according to user requirements, which can increase the system efficiency. The disadvantage is that the calculation overhead is large and the system delay is caused. Hybrid spectrum allocation integrates static and dynamic resource allocation methods to ensure that the system benefits are increased when the amount of calculation is low.
2. Classification by frequency band type: It is divided into two types: open spectrum sharing and licensed frequency spectrum sharing. Open spectrum sharing refers to accessing unlicensed bands with undifferentiated access to various communication users, such as the 2.4 GHz spectrum for industrial medical and scientific research. Authorized frequency band sharing means that for frequency bands already allocated to authorized users, secondary users can reasonably use spectrum resources without disturbing the normal communication of the primary user. In this way, the primary user's communication quality is prioritized, and secondary users need to dynamically exit or occupy the frequency band in response to changes in the environment.
3. Classification by access technology: It can be divided into Underlay mode and Overlay mode. The noise detected by the cognitive user as the primary user in the underlay mode exists, and the access to the frequency spectrum of the authorized user is controlled by controlling the cognitive user's transmit power and spread spectrum on the premise of guaranteeing the authorized user communication quality. This mode usually corresponds to the allocation. The interference temperature model in the algorithm. Overlay mode, cognitive users need to constantly detect the network environment, only the primary user does not occupy the frequency band, secondary users can access the opportunity, the primary user must also promptly perform spectrum switching.
4. Classification by collaboration: cooperative spectrum allocation and noncooperative spectrum allocation. In the cooperative allocation mode, cognitive users accessing a frequency band need to consider the influence and needs of other cognitive users, and maximize the overall benefit as a criterion for spectrum allocation. In a noncooperatively distributed model, each cognitive user of the frequency spectrum to be allocated can be considered "selfish" and allocate resources with the goal of maximizing self-

efficacy. This model reduces many users compared to a collaborative approach. The amount of information exchanged between the calculations.

5. Classification by network structure: According to whether or not the decision center is divided into centralized and distributed spectrum allocation. In the centralized network structure, each cognitive user with sensing function needs to send the detection result to the decision center. The center allocates the spectrum to achieve maximum system efficiency. The disadvantage is that the central computing has large overhead and the algorithm is complex. In the distributed distribution mode, each user transmits the information of the perception result. The distribution result is related to the criteria adopted by cognitive users. This entry studies the distributed spectrum allocation algorithm.

Key Applications

The spectrum allocation of cognitive radio has many common features with the spectrum allocation of other communication systems. However, due to the characteristics of cognitive radios borrowing the spectrum of authorized users by themselves, the spectrum allocation must meet certain special requirements. The distribution principle is as follows:

1. Guaranteeing flexibility: Cognitive radios are radios that can detect available spectrum resources and choose to borrow the authorized user spectrum to communicate. Therefore, the information of the available spectrum must be updated in real time, and once the authorized user resumes the use of a certain spectrum space, the cognitive user must exit the frequency band within a short time and select other frequency bands for communication (Liang 2008). In this way, the most important feature of spectrum allocation technology in cognitive radios that is different from other wireless communication spectrum
- allocations is to ensure flexibility. Any spectrum allocation technology for cognitive radios must have strong spectrum backoff and conversion capabilities, and due to the constant update of available spectrum information, the corresponding spectrum allocation algorithm must also meet the requirements of real-time performance.
2. Improving system performance: The primary purpose of spectrum allocation technology is to allocate the available spectrum space reasonably so that the system performance is improved or approached to an optimal state. Depending on the needs of different applications, the performance requirements of a certain cognitive system may also be different. For example, the goal is to minimize system interference, to increase the fairness of spectrum allocation, to maximize system throughput, and so on. According to different system application requirements, different algorithm objective functions are proposed to guide the design of the spectrum allocation algorithm.
3. Reducing signaling overhead and computational complexity: The design of the spectrum allocation algorithm will undoubtedly require a certain amount of algorithm signaling to transmit and occupy a certain amount of computing time. These can be seen as the overhead incurred by the allocation algorithm. Of course, it is hoped that the spectrum allocation function can be realized with relatively little overhead. Therefore, the spectrum allocation algorithm must be designed to take into account the complexity of control signaling between users and between the user and the central controller, distributed on the user or the central controller. The amount of algorithm calculation is also a problem that needs to be considered.

In short, the spectrum allocation of cognitive radio has a certain degree of universality and particularity and must be fully considered when designing spectrum allocation algorithms to meet the design principles listed above.

Future Directions

Game Theoretic Model: The cognitive radio spectrum allocation process can be abstracted as a game theoretic model. The cognitive user of each channel to be allocated in cognitive wireless network is a participant. The element in the behavior set of decision making is that the cognitive user selects the access channel from the list of available spectrum, and the benefit of network system after the allocation is finished. For example, spectrum efficiency is the utility of game process.

Interference Temperature Model: The interference temperature is similar to the noise temperature and is a comprehensive consideration of the interference generated by the user and its frequency band width. Cognitive users need to adjust the transmitting power or spectrum of the transmitter to not exceed the interference temperature of the main user by detecting the interference of the authorized user, and the power control plays a decisive role in the temperature interference model (Urkowitz 1967). According to the different interpretations of the signal and noise, it can be divided into the rational interference model and the generalized interference model.

Graph Theory Model: Spectrum allocation mathematical model is based on the corresponding interference and constraints. In the study of spectrum allocation in cognitive radio systems, the network topology of cognitive users is abstracted into a graph. Each vertex in the graph represents a wireless user and each edge represents a collision or interference between a pair of vertices.

Auction Bidding Model: Auction model requires an auctioneer to be responsible for allocating spectrum resources, so the model corresponds to the centralized spectrum allocation algorithm mentioned above, where each cognitive user, as a bidder, makes a price tag for the spectrum resources required. Unlike auction in the traditional economic sense, nodes cannot interfere with each other, such as neighbor users who are too close to each other cannot occupy the same frequency band.

Cross-References

- ▶ [Game Theory-Assisted Cooperative Spectrum Access](#)
- ▶ [MAC in Cognitive Radio Networks](#)
- ▶ [Spectrum Sensing In Cognitive Radio Networks](#)

References

- Cabric DS, Mishra M, Brodersen R (2004) Implementation issues in spectrum sensing for cognitive radios. In: Proceedings of the 38th Asilomar conference on signals, systems computers, Pacific Grove, pp 772–776
- Hossain E, Niyato D, Han Z (2009) Dynamic spectrum access in cognitive radio networks. Cambridge University Press, Cambridge, UK
- Liang YC, Zeng YH, Peh ECY, Anh TH (2008) Sensing-throughput tradeoff for cognitive radio networks. *IEEE Trans Wirel Commun* 7(4):1326–1337. <https://doi.org/10.1109/TWC.2008.060869>
- Mitola J (2001) Cognitive radio for flexible mobile multimedia communications, Mobile networks and applications. 6(5):435–441
- Urkowitz H (1967) Energy detection of unknown deterministic signals. *Proc IEEE* 55(4):523–531

Spectrum Sensing

- ▶ [Spectrum Sensing in Cognitive Radio Networks](#)

Spectrum Sensing in Cognitive Radio Networks

Yujie Tang
University of Waterloo, Waterloo, ON, Canada

Synonyms

[Cognitive radio](#); [Spectrum sensing](#); [Spectrum hole](#)

Definitions

Cognitive radio is a wireless communication system with intelligence, which senses its outside electromagnetic environment and learns from the surroundings and then adapts its internal states by changing certain operating parameters such as transmission power, carrier frequency, and modulation strategy in order to adapt to the changes of its environment.

Spectrum hole is a band of frequencies assigned to a primary user/licensed user but is not being utilized at a particular time and specific geographic location.

Spectrum sensing is the process of periodically monitoring a specific frequency band with the aim to identify the presence or absence of primary users/licensed users.

Historical Background

The useable electromagnetic radio spectrum which is a precious natural resource has been exclusively allocated to different wireless services by regulatory bodies. However, there have been several studies and reports over the years show that a large portion of licensed spectrum is actually underutilized in both temporal and spatial domains. On the other hand, the spectrum resource demands for services in the fifth-generation (5G) wireless communications, internet of things (IoT), and ad hoc vehicular network (VANET) are increasing dramatically. Therefore, there is a contradiction between the underutilized spectrum and increasing demand.

In order to solve the aforementioned challenge, cognitive radio (CR), regarded as a key technology, emerges as a novel form of wireless communications to improve spectrum efficiency. The concept of CR was first proposed by Joseph Mitola in 1999, and he defined a CR as "A radio that employs model based reasoning to achieve a specified level of competence in radio-related domains (Mitola 1999)." Mitola introduced a new language called the radio knowledge representation language (RKRL) which could be used by CR to enhance the flexibility of personal wireless

services. The idea of RKRL was expanded further in Mitola's doctoral dissertation, and this work was presented at the Royal Institute of Technology, Sweden, in May 2000 (Mitola 2000). In November 2002, the Federal Communications Commission (FCC) published a report prepared by the Spectrum Policy Task Force, aimed at improving the spectrum resource management in the United States (Kolodzy and Avoidance 2002), and the report was presented at a workshop focusing on CR, which was held in Washington, DC, in May 2003. This workshop was followed by a conference on CR held in Las Vegas, NV, in March 2004.

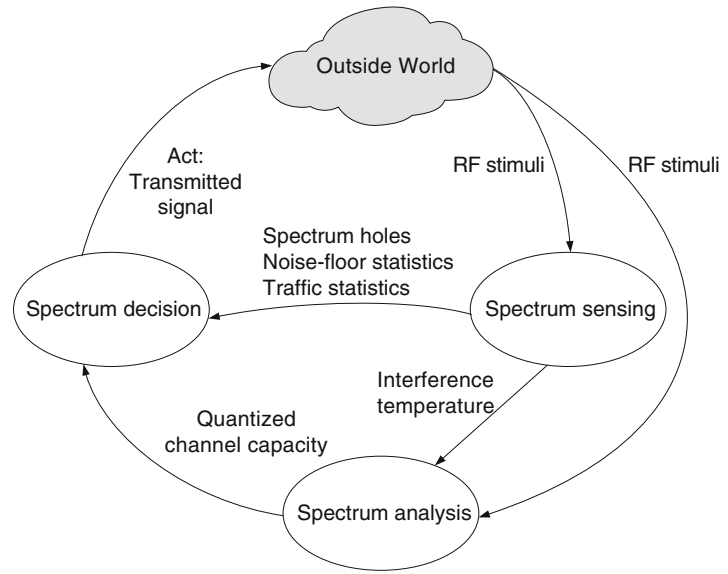
CR is built on a software-defined radio, and it is an intelligent wireless communication system that is aware of its environment (Haykin 2005). CR can learn from the radio environment and adapt to statistical variations in the input radio frequency (RF) stimuli by adjusting the transmitting parameters, such as power output, frequency, and modulation to ensure an optimized communications experience for users. Through interaction with the environmental RF stimuli, CR forms a cognitive cycle, which is demonstrated in Fig. 1.

In a cognitive radio network (CRN), there are two types of users: primary users (PUs) and secondary users (SUs). PUs are also referred to as licensed users, and SUs are referred to as unlicensed users. SUs probe for spectrum access opportunities through spectrum sensing to prevent harmful interference on the ongoing transmissions of PUs. Spectrum sensing plays an important role in CR. The aim of spectrum sensing is to detect a spectrum hole precisely in order to share the spectrum without harmful interference to other users. In order to guarantee the SUs can access the spectrum hole reliably and to utilize the band effectively without interfering with the PUs, various spectrum sensing techniques have been discussed in literatures.

Foundations

There are many ways to classify spectrum sensing schemes, and the most popular one is categorized by noncooperative spectrum

Spectrum Sensing in Cognitive Radio Networks, Fig. 1
Cognitive cycle



sensing and cooperative sensing. Noncooperative spectrum sensing is referred to as transmitter detection which mainly includes energy detection, matched filter detection, waveform-based sensing, and cyclostationary feature detection, while cooperative sensing is termed as receiver detection.

In noncooperative spectrum sensing, SUs detect the presence or absence of PU's signal, which can be regarded as a binary hypothesis testing problem. The received signal at the SU receiver can be expressed as

$$y(n) = \begin{cases} w(n), & H_0 \\ x(n) + w(n), & H_1 \end{cases} \quad (1)$$

where $w(n)$ denotes the additive white Gaussian noise (AWGN). The signal that comes from a PU transceiver and is received at the SU receiver is denoted as $x(n)$. H_1 and H_0 represent the presence and absence of the primary signal, respectively.

Energy detection (Digham et al. 2007) performs well if the noise power at the receiver is prior knowledge. However, the performance of the energy detector can be significantly degraded due to uncertainty in the noise power, which gives

rise to its unreliable nature. Thus, many recent research works focus on overcoming the noise uncertainty problem, i.e., cooperatively senses the noise energy to reduce the uncertainty to its minimal level.

A matched filter is a linear filter designed to maximize the output signal-to-noise ratio (SNR) of a given input signal, and it is known to be an optimal coherent detector in nature. Prior knowledge of the modulation type and carrier frequency of the PU is required for the matched filter detectors. Therefore, when the SU has prior knowledge not only about the various parameters of the physical implementation but also the physical structure of the primary signal, matched filter detection is applied. Otherwise, the correlation becomes poor, and the detection accuracy is degraded because of the unknown knowledge such as fading channel coefficients or a possible frequency offset.

Waveform-based detection is applicable to a system with known signal pattern which includes preambles, midambles, pilot patterns, spreading sequence, etc. A preamble is a known sequence transmitted before each burst, while a midamble is delivered in the middle of a burst or slot. Such kind of spectrum sensing method is also referred to as coherent sensing. Waveform-based detec-

tion has higher reliability and converges faster than energy detection. The performance of the waveform-based detection approach increases as the length of the known signal pattern increases.

Cyclostationary feature detection (Sutton et al. 2008) captures the features of received signals and identifies the noise energy of the modulated signal energy from the aperiodic characteristics of the noise signal. In other words, cyclostationary feature detection method deals with the inherent cyclostationary properties that cannot be found in any interference signal or stationary noise. More specially, the detection approach utilizes cyclic correlation periodicity in the received primary signal to identify the presence of PUs. Therefore, the cyclostationary feature detection method is a reliable method of spectrum sensing at low SNR as it possesses higher noise immunity than any other spectrum sensing schemes. But, on the other hand, this detection method requires prior knowledge of the signal parameters of primary signal, such as chip rate, channel code, and symbol rate, which has high computational costs (Gardner 1990).

In reality, since the primary signal suffers from multipath fading, the signal received at the SU receiver may not be sensed correctly. In addition, more accurate spectrum sensing performance takes longer observation time which will cause the incorrect spectrum utilization. Moreover, the hidden terminal problem that the SU receiver is blocked by a building or a wall can lead to incorrect spectrum sensing. To this end, cooperative spectrum sensing is emerged to overcome the aforementioned drawbacks in which SUs perform spectrum sensing in a coordinated and cooperative manner. In cooperative spectrum sensing, each SU independently implements local spectrum sensing and then reports either a binary decision (hard decision) or a test statistic (soft decision) to a combining user or fusion center. Then, the combining user or fusion center makes a decision on the presence or absence of the primary signal based on its received information. Through cooperation, SUs can share their sensing information for making a combined decision more accurate than the individual decisions. Therefore, the probability of misdetection or false

alarm is decreased by using cooperative spectrum sensing. The aim of cooperative spectrum sensing is to improve the sensing performance by exploiting the spatial diversity from spatially located SUs (Akyildiz et al. 2011). The performance improvement due to spatial diversity is called cooperative gain. Based on the methods used by SUs to share their sensing information, cooperative sensing schemes can further be classified into centralized sensing and distributed sensing.

In centralized sensing, a central controller collects sensing information from the SUs who have the sensing capability. Based on the obtained information, the central controller identifies the available spectrum and then broadcasts the spectrum information to all the SUs who are desiring for spectrum access opportunities. More specially, the sensing results with local decisions, either hard or soft, are gathered at the central controller through a control channel which is for reducing the probability of misdetection. However, this kind of sensing mechanism consumes lots of bandwidth. For distributed sensing, SUs share information among each other, but they make their own decisions of which part of the spectrum they can use. Distributed sensing has more advantages than centralized sensing, i.e., there is no need for a backbone infrastructure which reduces the sensing cost.

In spectrum sensing, another important issue is interference temperature detection, which measures the interference level observed at an SU receiver and is used to protect PUs from harmful interference due to SUs' opportunistic spectrum access. Although interference is regulated by the transmitters, it actually occurs at the receivers. Interference-based detection is one important method in CRN since an SU is allowed only if it does not degrade the quality of service (QoS) of a PU, which means an SU transmits only while the interference caused by its own transmission is below a specific threshold which is also referred to as the interference temperature limit. Each PU has an interference temperature limit which can guarantee a certain QoS of the PU. The SU will measure the interference environment and adjust its transmission such that the interference to the PU is not above the interference limits.

Thus, interference-based detection is an underlay scheme based on an estimation of the interference level at the primary receiver. An interference-based detection is discussed in Lin et al. (2013). Using the interference temperature parameter, two crucial controls can be defined as follows: (1) channel is declared to be occupied when the interference temperature is above the upper-bound threshold; (2) channel is declared empty or available for SUs while the interference temperature is below the lower-bound threshold.

Key Applications

Spectrum sensing technology is fundamental in the scenarios where the users would like to access the licensed bandwidth, i.e., the vehicles are using TV bands for communications in the vehicular ad hoc network.

Cross-References

- [Cooperative Cognitive Radio Networks](#)
- [Dynamic Spectrum Access Through Cognitive Radio Networking](#)
- [Spectrum Resource Allocation in Cognitive Radio Networks](#)

References

- Akyildiz IF, Lo BF, Balakrishnan R (2011) Cooperative spectrum sensing in cognitive radio networks: a survey. *IEEE Trans Mob Comput* 4:1874–4907
- Digham FF, Alouini MS, Simon MK (2007) On the energy detection of unknown signals over fading channels. *IEEE Trans Commun* 55:21–24
- Gardner WA (1990) Introduction to random processes: with applications to signals and systems. McGraw-Hill, New York
- Haykin S (2005) Cognitive radio: brain-empowered wireless communications. *IEEE J Sel Areas Commun* 23:201–220
- Kolodzy P, Avoidance I (2002) Spectrum policy task force. Federal Communications Commission, Washington, DC, Rep ET Docket 40:147–158
- Lin YE, Liu KH, Hsieh HY (2013) On using interference-aware spectrum sensing for dynamic spectrum access in cognitive radio networks. *IEEE Trans Mob Comput* 12:461–474
- Mitola J (1999) Cognitive radio: making software radios more personal. *IEEE Pers Commun* 6:13–18
- Mitola J (2000) An integrated agent architecture for software defined radio. PhD thesis, Royal Institute of Technology, Stockholm, Sweden
- Sutton PD, Nolan KE, Doyle LE (2008) Cyclostationary signatures in practical cognitive radio applications. *IEEE J Sel Areas Commun* 26:13–24

Spectrum Sensing with Power Recognition

Danyang Wang and Zan Li
Xidian University, Xi'an, China

Synonyms

[Signal detection](#); [Transmission power classification](#)

Definitions

Spectrum sensing with power recognition is the task of obtaining awareness about the existence of the primary user (PU) as well as the transmission power being used by the PU in a geographical area. The operators of spectrum sensing with power recognition are the secondary users (SUs) who have no licensed spectrum bands while seeking the opportunity to exploit the spectrum hole.

Historical Background

Spectrum sensing is the most important component to support the cognition cycle which has been first proposed by Joseph Mitola in 1999 (Mitola and Maguire 1999). The main function of spectrum sensing in this cognition cycle is to sense its surrounding environment (i.e., outside world). In 2005, a more concise cognition cycle was introduced by Simon Haykin, where the task of spectrum sensing was to detect the spectrum holes and measure the noise floor statistics

(Haykin 2005). Later on, IEEE Standards Association has made significant contributions in establishing the bridge between research results and widespread deployment of the spectrum sensing. More specifically, IEEE 1900.6 working group has started a standardization project on the topic of spectrum sensing in advanced radio systems. The baseline standard IEEE 1900.6 for spectrum sensing interfaces and data structures was released in 2011 (575 2011). Then the first amendment assigned as 1900.6a was released in 2014 (682 2014), and the latest amendment assigned as 1900.6b on the topic of the usage of spectrum sensing information to support spectrum databases was released in 2016 (741 2016). Moreover, IEEE 802.22 standard of wireless regional area networks (WRAN) in TV bands was released in 2011, where spectrum sensing was a key function of the cognitive plane (ieee 2012). With the release of IEEE 802.22 series, LTE, and LTE-A standards, the PU could operate under more than one discrete transmission power. Thus, spectrum sensing with power recognition that can detect the presence and transmission power of the PU has attracted more attention recently.

Foundations

With cognitive radio being used in a number of applications, the area of spectrum sensing with power recognition to effectively detect the presence and the transmission power of the PU has become increasingly important. The basic ability of spectrum sensing with power recognition is to keep monitoring the spectrum band to ensure that the SU cannot cause harmful interference to PUs. When performing spectrum sensing with power recognition, the SU may work with two kinds of operation types:

- **Periodic sensing:** When a single RF chain is utilized to process both the sensing and data transmission in the single radio architecture, only a specific time slot can be allocated for spectrum sensing and power recognition. In

this case, the spectrum efficiency is limited, since a part of the available time slot is used to perform spectrum sensing instead of data transmission (Zhang et al. 2014b).

- **Continuous monitoring:** The capability that continuously senses the spectrum occupancy and identifies the transmission power is very necessary for the SU in practical cognitive radio systems. The SU is allowed to exploit the available spectrum on a noninterference basis to the PU. When the PU returns to the spectrum, the SU needs to detect the PU's activity immediately by leveraging continuous monitoring (Zhang et al. 2014a). The continuous monitoring can be realized in the dual-radio architecture, where one radio chain is allocated to data transmission, while the other chain is dedicated to spectrum monitoring.

Technically, both periodic sensing and continuous monitoring rely on the SU which can be able to detect very weak signals transmitted by the PU and identify the transmission power of the PU. This sophisticated processing procedure is usually completed by excavating a specific feature of the PU's transmitted signal. On theoretical aspect, detecting the presence and identifying the transmission power of the PU can be described by the multiple hypotheses test problem. In this framework, the SU declares a specific hypothesis is true only if the measurement of signal features is located in the predetermined decision region. A number of features can be leveraged to calculate the measurement, which yields to various spectrum sensing with power recognition schemes.

Matched filter detector-based scheme is known as the optimal coherent detector under additive white Gaussian noise (AWGN) channels. In this scheme, the test statistics is calculated by employing cross-correlation between the known sequence and the received signal (Zhang et al. 2015). Then, the occupancy of the spectrum is detected by examining the test statistics with predetermined threshold. Moreover, the transmission power used by the PU is identified by utilizing minimum Bayes risk criterion. Although matched filter-based scheme can provide a superior performance, it requires the

SU to be aware of prior information about the PU. Practically, it is not always easy to obtain the prior information of the PU. In this case, the sensing and recognition performance of matched filter-based scheme degrades severely.

Energy detector-based scheme is the most common way of spectrum sensing with power recognition due to low computational complexity. Moreover, the energy detector-based scheme requires no prior knowledge of the PU's signal, which makes it more generic (Gao et al. 2015). In the energy detector-based scheme, the absolute square of the received signal is calculated and averaged over the sensing slot to obtain the test statistics. The activity status as well as the transmission power of the PU is detected by comparing the test statistics with predetermined decision thresholds. However, the energy detector-based scheme makes poor performance under low signal-to-noise ratio. In addition, the noise power is required to calculate decision thresholds, such that a lower estimation error on the noise power may result in a severe degradation on the sensing and recognition performance.

When the SU is equipped with multiple antennas, a promising spectrum sensing with power recognition scheme was proposed Li et al. (2016), where the maximum eigenvalue of the received sample covariance matrix is used to identify the status of PU. To calculate the decision regions, it is very necessary to derive the probability density function (PDF) of the maximum eigenvalue firstly. The exact PDF of the maximum eigenvalue can be derived by leveraging the random matrix theory. However, the exact PDF of the maximum eigenvalue is extremely complicated. Thus, it is hard to utilize the exact PDF to calculate the decision regions. In practice, non-asymptotic Gaussian approximation and gamma approximation are usually used to obtain the PDF of the maximum eigenvalue by fitting the first- and second-order moments of the maximum eigenvalue. Therefore, the closed-form expression of the decision regions as well as the closed-form performance analysis can be derived concisely. The eigenvalue-based spectrum sensing with power recognition scheme requires no prior

information of the PU's signal, which makes it a blind spectrum sensing scheme.

When the transmitted signals are non-Gaussian process, the high-order cumulant-based spectrum sensing with power recognition scheme can efficiently detect the presence of the PU and recognize the transmission power of the PU by leveraging the high-order statistical feature of the received signal (Wang et al. 2018). It is well known that cumulants higher than second order are zero for a Gaussian random process. Hence, the high-order cumulant-based spectrum sensing with power recognition scheme can eliminate the effect of Gaussian noise on the sensing and recognition performance. Additionally, to achieve more accurate sensing and recognition performance, hybrid multiple high-order cumulant-based spectrum sensing and power recognition has been proposed, where the rich statistical information of the received signal is excavated.

To illustrate the performance of spectrum sensing with power recognition scheme, three probabilities are commonly explored: probability of false alarm, probability of detection, and probability of recognition. Probability of false alarm is the probability of an incorrect decision, where the SU claims the monitored spectrum is occupied when the PU is actually inactive. Probability of detection is the probability of detecting a signal on the monitored spectrum band when the PU is truly active. Furthermore, probability of recognition describes the capability that the SU correctly identifies the transmission power being used by the PU. Generally, probability of false alarm should be kept as small as possible, in order to ensure the opportunity for reusing the idle spectrum, while the probability of detection and probability of recognition should be achieved as high as possible to prevent the PU from being harmfully interfered.

Key Applications

Spectrum sensing with power recognition can be used in LTE-U systems, where the cellular network shared unlicensed spectrum with Wi-

Fi system. Specifically, the spectrum sensing with power recognition is utilized to find available unlicensed spectrum band, which can be exploited by the cellular network to offer higher capacity without affecting the performance of Wi-Fi systems.

Cross-References

- [Cognitive Radio Networks](#)
- [Cooperative Cognitive Radio Networks](#)
- [Spectrum Sensing in Cognitive Radio Networks](#)

References

- (2011) IEEE standard for spectrum sensing interfaces and data structures for dynamic spectrum access and other advanced radio communication systems. IEEE Std 19006-2011, pp 1–168. <https://doi.org/10.1109/IEEESTD.2011.5756728>
- (2012) Ieee standard for information technology-part 22: cognitive wireless ran medium access control (MAC) and physical layer (PHY) specifications: policies and procedures for operation in the tv bands. IEEE Std 80222-2011, pp 1–15
- (2014) Ieee standard for spectrum sensing interfaces and data structures for dynamic spectrum access and other advanced radio communication systems – amendment 1: procedures, protocols, and data archive enhanced interfaces. IEEE Std 19006a-2014 (Amendment to IEEE Std 19006-2011), pp 1–52. <https://doi.org/10.1109/IEEESTD.2014.6823063>
- (2016) Ieee standard for spectrum sensing interfaces and data structures for dynamic spectrum access and other advanced radio communication systems – corrigendum 1. IEEE Std 19006-2011/Cor 1-2015 (Corrigendum to IEEE Std 19006-2011), pp 1–15. <https://doi.org/10.1109/IEEESTD.2016.7419198>
- Gao F, Li J, Jiang T, Chen W (2015) Sensing and recognition when primary user has multiple transmit power levels. IEEE Trans Signal Process 63(10):2704–2717. <https://doi.org/10.1109/TSP.2015.2415751>
- Haykin S (2005) Cognitive radio: brain-empowered wireless communications. IEEE J Sel Areas Commun 23(2):201–220
- Li Z, Wang D, Qi P, Hao B (2016) Maximum-eigenvalue-based sensing and power recognition for multiantenna cognitive radio system. IEEE Trans Veh Technol 65(10):8218–8229. <https://doi.org/10.1109/TVT.2015.2511783>
- Mitola J, Maguire GQ (1999) Cognitive radio: making software radios more personal. IEEE Pers Commun 6(4):13–18. <https://doi.org/10.1109/98.788210>
- Wang D, Zhang N, Li Z, Gao F, Shen X (2018) Leveraging high order cumulants for spectrum sensing and power recognition in cognitive radio networks. IEEE Trans Wirel Commun 17(2):1298–1310. <https://doi.org/10.1109/TWC.2017.2777488>
- Zhang N, Cheng N, Lu N, Zhou H, Mark JW, Shen XS (2014a) Risk-aware cooperative spectrum access for multi-channel cognitive radio networks. IEEE J Sel Areas Commun 32(3):516–527. <https://doi.org/10.1109/JSAC.2014.1403004>
- Zhang N, Liang H, Cheng N, Tang Y, Mark JW, Shen XS (2014b) Dynamic spectrum access in multi-channel cognitive radio networks. IEEE J Sel Areas Commun 32(11):2053–2064. <https://doi.org/10.1109/JSAC.2014.141109>
- Zhang X, Gao F, Chai R, Jiang T (2015) Matched filter based spectrum sensing when primary user has multiple power levels. China Commun 12(2):21–31. <https://doi.org/10.1109/CC.2015.7084399>

Sponsored Content

- [Game Theoretic Modeling of Zero-Rating Internet Provisioning](#)

Sponsored Data

- [Game Theoretic Modeling of Zero-Rating Internet Provisioning](#)

Spontaneous Networking

- [Evolution of Vehicular Networks in the Transition from 3G to 4G Systems](#)

Spoofing Attack

- [Pilot Spoofing Attack](#)

Spread Spectrum

- ▶ [Wireless Jamming Attack](#)

Statistical Channel State Information

- ▶ [Massive MIMO BDMA Transmission](#)

Stochastic Geometry

- ▶ [Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes](#)

Study on Machine-Learning Algorithms in LTE Networks

- ▶ [Survey on Machine-Learning Techniques in LTE Networks](#)

Subcarrier-Index Modulation Aided Orthogonal Frequency Division Multiplexing

- ▶ [Index Modulation for OFDM](#)

Successive Interference Cancellation

- ▶ [Non-orthogonal Multiple Access \(NOMA\)](#)

Superposition Coding

- ▶ [Non-orthogonal Multiple Access \(NOMA\)](#)

Survey on Machine-Learning Techniques in LTE Networks

Fangmin Xu¹, Jia Hou², and Fan Yang¹

¹Beijing University of Posts and Telecommunications, Beijing, China

²Department of Communications Engineering, Soochow University, Suzhou, Jiangsu, China

Synonyms

[Machine-Learning \(ML\) in 4G networks](#); [Study on Machine-Learning algorithms in LTE Networks](#)

Definitions

Machine-learning (ML) techniques in LTE networks mean to apply machine-learning algorithms in LTE network to improve the network performance.

History Background

Learning is a kind of important intelligent behavior that human beings have. At present, computers have also acquired this ability initially, and this is machine learning (ML). A classic definition is machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. In short, machine learning refers to the acquisition of new experiences and knowledge through the inherent regularity of computer learning data to improve the intelligence of computers and enable computers to make decisions like human beings.

ML is an important subject that gives machine intelligence and solves many problems. ML techniques can be applied to different areas of LTE. And history of machine learning can be traced back to the twentieth century.

- In 1943, Warren McCulloch and Walter Pitts proposed the neural network hierarchy model and established a computational model theory of neural network, which laid the foundation for the development of machine learning.
- In 1950, the “Father of Artificial Intelligence” Turing developed the famous “Turing Test,” making artificial intelligence an important research topic in the field of computer science.
- Till the summer of 1980, the first international symposium on machine learning, held at Carnegie Mellon University in the United States, marked the rise of machine learning in the world. In the 1980s, researches about neural network increase gradually, which includes backpropagation algorithm (BP) (Hopfield 1982), convolutional neural network (CNN), and so on. These are based on knowledge from samples.
- In the 1990s, many statistical learning models have attracted the researchers’ attention, such as support vector machine, kernel method, and so on. However, research on this field began from the 1960s. Vapnik proposed the concept of support vector in 1963 (Vapnik 1998). Since the 1990s, the superiority of statistical learning is shown in text classification application.
- In the 2000s, the problem of shallow learning is gradually exposed, because the limited sample and calculation unit led to limited expression ability of complex functions and data. The learning ability is not strong, which can only extract the primary characteristics. Thus, in 2006, Geoffrey Hinton published an article (Hinton and Salakhutdinov 2006), which proposed a deep learning model. The introduction of this model opens up a new era of deep neural network machine learning. Deep learning in recent years has made impressive progress in many fields and launched a number of successful commercial applications, such as Google Translate, Apple Siri voice tools, Microsoft Cortana, and especially the Google AlphaGo which uses Monte Carlo tree search algorithm and artificial neural networks to defeat the European Go champion Fan Hui (Fu 2016).

Since LTE launched in 2004, few years later, research on machine-learning techniques in LTE networks becomes hot topic. Till now, there are still many researchers focus on this field.

Foundations

Machine learning has becoming a useful tool to solve complex problems in many fields. Meanwhile, many functions and implementations in LTE network require the assistance of intelligent technologies.

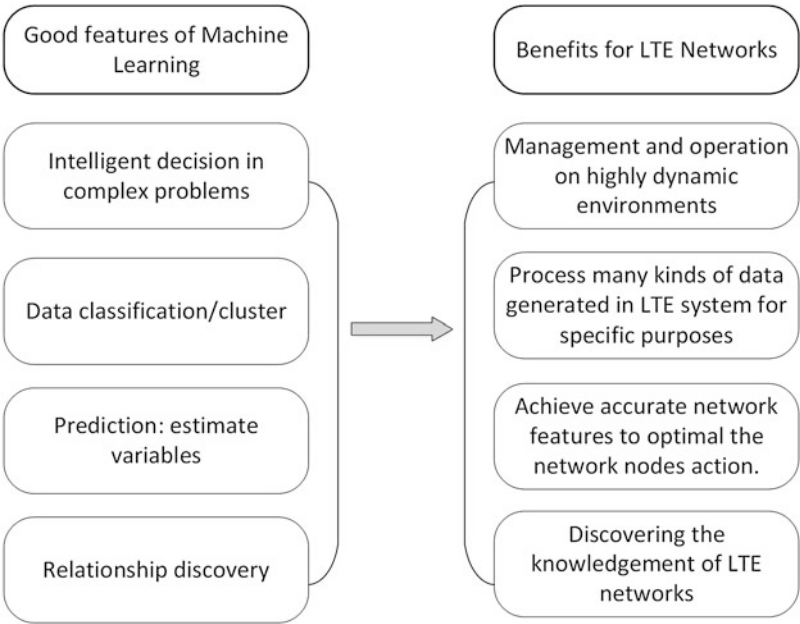
Traditionally, quasi-static predefined set of rules (or policies) are stored in the LTE network devices. For example, the handoff process will be triggered once the RSRP of serving cell below a certain level for a period of time without considering other factors (such as network load and interference states) in network. The rule-based system is simple and convenient to implement. However, they cannot be changed with respect to the dynamically changing wireless environment; therefore the real performance could not reach the optimal ideal performance. For instance, the propagation channels, spectrum availability, topology, and interference are uncertain factors that are too complex to solve by static algorithm. Hence, the idea of incorporate artificial intelligence into the LTE system design arose gradually. As a way to realize artificial intelligence, machine learning uses various algorithms to train the massive dataset, learn from it, and then make decisions and predict events in the real world, so that optimal or near-optimal performance can be achieved. Figure 1 shows the good features of machine learning which can benefit the LTE networks.

• Intelligent Decision in Complex Problem

Decision-making is selecting a course of action from among available alternatives. There exist many complex decision problems in LTE system. The mobility of mobile devices and the dynamic nature of wireless channel result in the highly dynamic

Survey on Machine-Learning Techniques in LTE Networks, Fig. 1

Good features of machine learning that can benefit LTE networks application

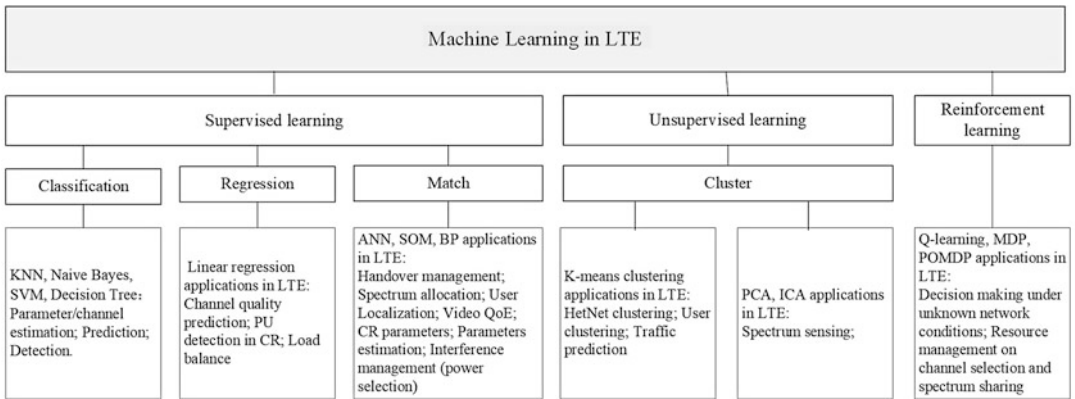


environments and complicated management of spectrum access, radio resource, and so on. The highly complex nature of aforementioned problems in LTE system often means that inventing specific algorithms that will solve them perfectly every time is impractical, if not impossible. For example, the access control mechanism in HetNet directly influences the capacity of the cell. The typical factors that need to be considered in access control algorithms include the number of user equipments (UEs) with various QoS classes, channel quality indicator (CQI), resource allocation, and QoS requirements of different UEs. To achieve optimal performance, smart access control algorithm should automatically adjust according to the influent factors.

- **Data Classification/Cluster**
Classification and clustering are both methods that divide the data into classes based on similarity in the data. The difference is that classification is a kind of supervised learning with labeled training data, for example, spam/non-spam or fraud/non-fraud. The decision being modeled is to assign labels to new unlabeled pieces of data. This can be thought of as a dis-

crimination problem, modeling the differences or similarities between groups. On the other side, as unsupervised learning, data in clustering is not labeled but can be divided into groups based on similarity and other measures of natural structure in the data. In practical, many kinds of data generated in LTE system need to be classified or grouped for specific purposes, such as channel measurement report used for state feedback and description.

- **Prediction**
Prediction is estimating the value of a variable based on the value of another variables. The stronger the relationship between the variables, the more accurate the prediction. Regression is one of the prediction techniques based on correlation. Typical examples that are easy to understand are time series data like the price of a stock over time, the decision being modeled is what value to predict for new unpredicted data. The prediction of future data traffic requirement and mobility pattern based on historic data is very helpful for the operation of LTE system. For instance, the accurate prediction of future traffic and spectrum requirement would be helpful for



Survey on Machine-Learning Techniques in LTE Networks, Fig. 2 Machine-learning algorithms applicable to LTE

the intelligent spectrum allocation between multiple radio access technologies (RATs) such as GSM, UMTS, and LTE.

- **Relationship Discovery**
Data is used as the basis for the extraction of propositional rules. Such rules may, but are, typically not directed, meaning that the methods discover statistically supportable relationships between attributes in the data, not necessarily involving something that is being predicted. An example is the discovery of the relationship between the purchase of beer and diapers. In several operation processes of LTE system, there are many features that may influent the performance. However, generally the relationship between the features and the performance is not clear explicitly. With the help of machine-learning techniques, the inside relationship would be discovered and be used for optimizing the system performance.

Figure 2 shows the specific machine-learning algorithms applicable to long-term evolution.

Key Applications

LTE focuses on both performance requirements and efficiency requirements. Machine-learning techniques can be applied to both two areas to

improve the network performance and bring better user experience (UE). Machine learning has becoming a useful tool to solve complex problems in many fields. Meanwhile, many functions and implementations in LTE network require the assistance of intelligent technologies.

The performance requirement of LTE specify the peak data rate, mobility, latency, and so on. To satisfy these requirements, LTE developed many techniques which include RRM, CSI, channel prediction, traffic requirement prediction, and so on. The machine-learning algorithm which could bring network efficiency and intelligence is introduced to optimize the performance of the LTE system.

Table 1 summarizes the detailed application of machine learning in LTE network which shows the application areas, using machine-learning tools, and the purpose. It could help to sort out the skeleton of application of ML in LTE and make the structure clearer.

- **Resource Management**
Resource management in LTE is to provide high-quality service under the limited resources. The aim is to flexibly allocate and dynamically adjust the available resources of wireless network to maximize the utilization of resources. It includes power control, channel allocation, handover, scheduling, access control, load control, end-to-end QoS, and adaptive coding modulation. ML algorithms

Survey on Machine-Learning Techniques in LTE Networks, Table 1 Application of machine learning in LTE network

Applications		ML tools	Purpose
Resource management (performance)	Radio resource management	Q-learning, random neural network (RNN)	Adjust radio parameters to realize power control and decrease interference
	Access control/load balance	Feedforward neural network (FFNN), Q-learning	Decide whether to access certain eNodeB and decrease collision
	Handover	Self-organizing map (SOM), neural network, Q-learning	Learn the optimal handover parameters and decide to handover or not
	Channel state information (CSI)	Q-learning, neural network, support vector machine (SVM)	Optimize channel quality information (CQI) report way and even predict CQI
	Interference	Q-learning	Adjust power control parameters based on the environment to decrease interference
Prediction (performance)	Traffic load	K-means (KM), k-nearest neighbor (kNN)	Predict traffic and mobile data usage
	Channel quality	Random forest, neural network, regression tree	Predict channel quality and connectivity
Spectrum efficiency (efficiency)	Cognitive radio (CR)	Q-learning, artificial neural network (ANN), random neural network (RNN)	Determine the CR parameters by learning algorithms
Energy efficiency (efficiency)	Mobility terminals	Neural network, support vector regression (SVR)	Predict whether there will be control information to reduce the energy consumption of receiving control information
	eNodeB	Online Q-learning	Automatically switch on/off eNodeB
Security (efficiency)	SPAM/malware	Adaboost, Naive Bayes (NB), SVM	SPAM, malware detection

such as Q-learning, Neural Network could be used in Resource Management could increase the level of resource utilization (Kumar et al. 2013).

- Prediction

It is essential for people to employ prediction techniques in many areas such as mobile network traffic, throughput, and channel quality, so that users can get an increasing feel of satisfaction and the network runs more stable. To achieve these tasks, ML techniques such as k-nearest neighbor (kNN), k-means (KM), neural network, random forest, and other algorithms are used. By using these ML algorithms, the features of the network

could be determined, and then the prediction is made based on the features (Yue et al. 2017). Compared to tradition ways, the accuracy and efficiency of prediction are highly improved.

Efficiency requirements of LTE include the spectrum efficiency, energy efficiency, and even security management. This section will provide a detailed introduction of the machine-learning techniques using these techniques of enhancing system efficiency.

- Spectrum Efficiency

Since the increasing data rate in the radio network, more bandwidth is in need. However,

the spectrum is limited which requires more researches on new radio spectrum management pattern. Thus, many spectrum management patterns such as cognitive radio (CR) and other patterns based on prediction become more and more popular. However, the efficiency still needs to be improved to provide better service, so ML algorithms are introduced to this area. And usually, neural network algorithms could be used here, such as artificial neural network (ANN) and random neural network (RNN). ANN has been widely used in predicting in CR, and RNN has been proposed to achieve better generalization and to speed up the cognition process in LTE (Adeel et al. 2014). The results show that the network performance is greatly improved after the machine-learning algorithm is applied.

- **Energy Efficiency**

Due to the upsurge of traffic requirement as well as global energy resources shortage, energy saving in telecommunication systems attracts researchers' attention. Nowadays, energy saving could be realized by designing efficiency hardware, using renewable energy, and adopting efficiency resources allocation algorithm, base station sleeping mode techniques, and energy saving on mobile terminals. Since energy saving on mobile devices and base stations make no change on current network architecture and easy for implementation, so energy saving is usually realized by these two ways (Sue et al. 2016; Kong and Panaitopol 2013). ML algorithms could also be used here and usually help to optimize the network parameters to continuously adapt to the changing network traffic in deciding which action to take to minimize energy consumption.

- **Security**

Since the amount of data transferred in communication system is rising and the network environment is becoming complex, security in LTE network faces more and more challenge. Based on ML algorithms such as Adaboost, Naive Bayes (NB), and Support Vector Machine (SVM), the general action mode of different threats types could be

understood (Cao et al. 2017). This has been proved to help increase the accuracy of SPAM or malware detection in LTE network.

Future Directions

From the development of application of machine learning in LTE Networks, it appears that, in the future, emphasis will be given on three main directions. Firstly, future may focus on real-time decision instead of feature-based predictive applications. Thus, network nodes could make decisions based on real-time network conditions. Moreover, in the real world, human errors in data entry process, incorrect sensor reading, and software bug in data processing pipeline all could lead to missing values. Thus, secondly, the processing scheme for data sets with missing value may be paid more attention. Finally, overfitting problem occurs when the hypothesis becomes overly restrictive to obtain a consistent hypothesis. It is a general problem in machine learning, and it needs to be solved to obtain the nature rules of the data sets.

Cross-References

- [Energy Efficiency of Cellular Networks](#)
- [Reinforcement Learning-Based Wireless Communications Against Jamming and Interference](#)

References

- Adeel A, Larijani H, Ahmadiania A (2014) Performance analysis of random neural networks in lte-ul of a cognitive radio system. In: 2014 1st international workshop on cognitive cellular systems (CCS), pp 1–5. <https://doi.org/10.1109/CCS.2014.6933800>
- Cao J, Drabeck L, He R (2017) Statistical network behavior based threat detection. In: Computer communications workshops
- Fu MC (2016) Alphago and Monte Carlo tree search: The simulation optimization perspective. In: 2016 winter simulation conference (WSC), pp 659–670. <https://doi.org/10.1109/WSC.2016.7822130>
- Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507

- Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 79(8):2554–2558
- Kong PY, Panaitopol D (2013) Reinforcement learning approach to dynamic activation of base station resources in wireless networks. In: 2013 IEEE 24th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), pp 3264–3268. <https://doi.org/10.1109/PIMRC.2013.6666710>
- Kumar D, Kanagaraj NN, Srilakshmi R (2013) Harmonized q-learning for radio resource management in LTE based networks. In: 2013 proceedings of ITU kaleidoscope: building sustainable communities, pp 1–8
- Sue JA, Hasholzner R, Brendel J, Kleinstueber M, Teich J (2016) A binary time series model of LTE scheduling for machine learning prediction. In: 2016 IEEE 1st international workshops on foundations and applications of self* systems (FAS*W), pp 269–270. <https://doi.org/10.1109/FAS-W.2016.64>
- Vapnik V (1998) Statistical learning theory. Wiley, New York
- Yue C, Jin R, Suh K, Qin Y, Wang B, Wei W (2017) Linkforecast: cellular link bandwidth prediction in LTE networks. *IEEE Trans Mob Comput* PP(99):1–1. <https://doi.org/10.1109/TMC.2017.2756937>

Symbol Synchronization of Nanomachines

► Synchronization of Nanomachines

Synchronization of Nanomachines

Lin Lin¹ and Hao Yan²

¹Tongji University, Shanghai, China

²Shanghai Jiao Tong University, Shanghai, China

Synonyms

Clock synchronization of nanomachines; Symbol synchronization of nanomachines; Time synchronization of nanomachines

Definitions

Synchronization refers to a process of establishing a relation among nanomachines on time scale, such that the nanomachines can synchronize their behaviors/actions coordinately. Clock/time synchronization of nanomachines refers to setting the clock of different nanomachines to realize the same time. Symbol synchronization of nanomachines refers to aligning the symbol interval, especially the start of the symbol, between the transmitter nanomachine and the receiver nanomachine.

Historical Background

Synchronization is an important and basic issue in molecular communication and its related applications. Different from using conventional electric or electromagnetic signal to carry information, molecular communication uses molecules to realize information carrying. This fundamental difference and its new application scenarios bring the synchronization issue many new problems and new challenges.

There are several kinds of concept of the synchronization among nanomachines: synchronization of oscillation, time/clock synchronization, and symbol synchronization.

For *synchronization of oscillation*, Abadal and Akyildiz have modeled a bacteria quorum-sensing mechanism, which forms the oscillation pattern for each nanomachine, and at the same time, these oscillation patterns are synchronized automatically under the quorum-sensing mechanism (Abadal and Akyildiz 2011). Different from (Abadal and Akyildiz 2011) which uses autoinducer molecules, Moore and Nakano have proposed another mechanism using inhibitory molecules to realize the similar oscillation and synchronization (Moore and Nakano 2013). Lin et al. have introduced an oscillation and synchronization model with one nanomachine having a repressing process module (Lin et al. 2016a).

For *time/clock synchronization*, the concept of the clock of the nanomachine has been mentioned in Nakano et al. (2014). Similar to electronic microprocessor, the clock of the nanomachine keeps track of time. A molecular phase-locked loop (PLL) has also been proposed in Lo et al. (2014) for regulating the clock. A typical clock synchronization scheme for diffusion-based molecular communication has been proposed in Lin et al. (2016b). A two-way message exchange mechanism is proposed for a Gaussian propagation delay model. The clock offset and clock skew are synchronized by maximum likelihood (ML) estimation. Later on, a series of variations are proposed. Different from previous literature, a high-efficiency blind synchronization scheme, which uses only one symbol transmission, is proposed in Luo et al. (2018). Least square method and single-input multiple-output (SIMO) design are proposed to estimate the propagation delay and further achieve the clock synchronization.

For *symbol synchronization*, Shahmohammadian et al. have proposed a blind ML criterion for estimating the channel propagation delay to achieve the symbol synchronization (ShahMohammadian et al. 2013). Jamali et al. have considered some practical challenges where the transmitter may not be equipped with internal clock and may not be able to emit molecules with a fixed frequency (Jamali et al. 2017). Synchronization-detection frameworks employing two types of molecules have been developed. Here the symbol synchronization refers to the estimation of the start of a symbol interval.

Foundations

The concept of the synchronization of nanomachines here includes three kinds of synchronization, i.e., synchronization of oscillation, time/clock synchronization, and symbol synchronization.

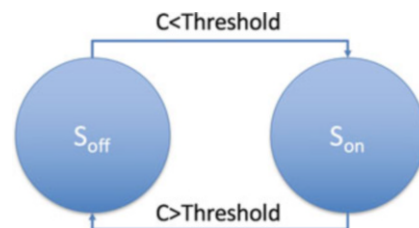
Synchronization of oscillation: This kind of synchronization originates from the quorum-sensing

mechanism in biology. In biology, quorum sensing is used by bacteria to regulate specific gene expression. In the field of molecular communication, nanomachines can adopt this mechanism or its variants to form a synchronized oscillation. An example is given below:

Example 1 It is assumed that the nanomachine can control the release of molecules and can sense the concentration of the molecules in its close environment. A loop can be established based on the following rules:

- If the sensed concentration is below certain threshold, the nanomachine will release a given amount of molecules.
- If the sensed concentration is above that threshold, the nanomachine will not release any molecules.
- The molecules diffuse away after releasing, so the concentration decreases. Once the concentration falls below the threshold, the above process will take place again.

The whole process can be represented by a finite-state machine (FSM) as shown in Fig. 1. “S_{off}” represents the nonactivated state, i.e., the nanomachine does not release molecules, while “S_{on}” represents activated state, i.e., the nanomachine releases molecules. If the sensed concentration reaches above the threshold, the state will change from S_{on} to S_{off}. If the sensed concentration falls below the threshold, the state will



Synchronization of Nanomachines, Fig. 1 State diagram for the FSM in the synchronization of oscillation. “S_{off}” and “S_{on}” represent the nonactivated and activated state. The state changes depending on the comparison result of the sensed concentration and the threshold

change from S_{off} to S_{on} , and the nanomachine will release an amount of molecules.

If there exist two or more nanomachines with the same working rules and use the same kind of molecules and they are close to each other under which condition the molecules released by one nanomachine can diffuse and arrive at other nanomachines, then eventually, the oscillations of these nanomachines will be synchronized.

Time/clock synchronization: Time/clock synchronization is essential for many applications. For example, the engineered nanomachines need a common clock so that they can release the antibody molecules at the same time to destroy tumors in the human body. In the data gathering application of the Internet of Nano Things (IoNTs), clock synchronization is the basis for interpreting the sensed data.

For different molecular communication channels and systems, different time/clock synchronization methods will be used. Two examples are provided below:

Example 2 The reference nanomachine sends its own clock reading to another nanomachine, which updates its clock once the clock reading is received using the following equation

$$T_{\text{rx}} = T_{\text{tx}} + t_{\text{delay}}, \quad (1)$$

where T_{tx} is the molecule releasing time instant based on the clock of the transmitter. The M -ary molecular shift keying (MoSK) is used to embed the clock reading into the molecule (Kim and Chae 2013). Here the time duration for synthesizing molecule is ignored. The key for the clock synchronization is to calculate the propagation delay t_{delay} .

The channel impulse response (CIR) in a 3D free diffusion environment without drift can be expressed as

$$N(t) = \frac{V_R Q}{(4\pi D t)^{3/2}} \exp\left(-\frac{d^2}{4Dt}\right), \quad (2)$$

where $N(t)$ represents the average number of observed molecules at time t , d is the distance from the transmitter to the receiving antenna, D is the diffusion coefficient, V_R is the volume of the spherical receiving antenna, Q is the number of released molecules.

t_{delay} is defined as the peak concentration time. In the absence of noise, t_{delay} can be derived from (2) as

$$t_{\text{delay}} = \frac{d^2}{6D}. \quad (3)$$

In case that d and D are known, t_{delay} can be directly calculated by (3). In case that d and D are unknown, the sampled concentration signals are used to estimate d and D which are further used to calculate t_{delay} by (3). The detailed processes are as follows. Because the nanomachines are not synchronized, so (2) in the 3D free diffusion channel is modified as

$$N(t) = \frac{V_R Q}{(4\pi D(t - t_0))^{3/2}} \exp\left(-\frac{d^2}{4D(t - t_0)}\right), \quad (4)$$

where t is a time instant based on the receiver's system clock. The time instant for molecule releasing based on the receiver's system clock is t_0 . Because there are limited unknowns in (4), so if enough samples are taken, then the unknowns can be calculated. And further t_{delay} can be estimated. According to (1), the receiver calculates T_{rx} and updates its clock value with it.

Example 3 In this example, the propagation delay (i.e., molecule arrival time) is not constant but a random variable because the channel is an inverse Gaussian channel in this case. Therefore, statistical methods can be used to estimate the clock offset (and clock skew if needed) between the nanomachines, so that they can update their clocks to achieve the clock synchronization.

Assuming that in a 1D drift channel, the transmitter sends its clock reading $T_{\text{tx},i}$ immediately into the channel with a molecule. MoSK is used to embed the clock reading into the molecule. Once receiving the molecule, the receiver records

its own clock reading $T_{rx,i}$ which is

$$T_{rx,i} = T_{tx,i} + t_i + \theta, \quad (5)$$

where t_i is propagation delay for the i th message transmission, which is a random variable. θ is the clock offset between the two nanomachines, which needs to be estimated.

It is known that the propagation delay t_i of the molecule in a 1D drift channel follows an inverse Gaussian distribution (Srinivas et al. 2012). The probability density function (PDF) can be expressed as

$$f(t_i; \mu, \lambda) = \left(\frac{\lambda}{2\pi t_i^3} \right)^{1/2} \exp \frac{-\lambda(t_i - \mu)^2}{2\mu^2 t_i}, \quad (6)$$

where μ and λ are related to the channel parameters which are assumed to be known by the nanomachine. After N message transmissions in total shown in Fig. 2, $\{T_{tx,i}, T_{rx,i}\}_{i=1}^N$ is obtained. Then based on the recorded time stamps $\{T_{tx,i}, T_{rx,i}\}_{i=1}^N$ and the PDF, one can use the ML estimation to estimate the clock offset θ , which can be expressed by

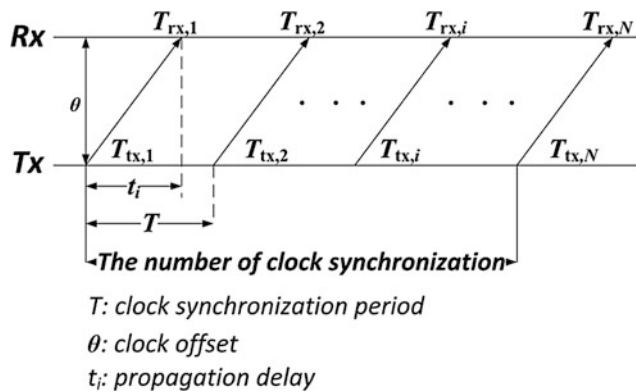
$$\hat{\theta} = \arg \max_{\theta} [\ln L(\theta; \{T_{tx,i}, T_{rx,i}\}_{i=1}^N)]. \quad (7)$$

where $L(\cdot)$ is likelihood function.

The ML method in Example 3 is suitable for the case where the PDF of the propagation delay and the system parameters in the PDF are known. If the PDF of the propagation delay is unknown but the CIR is known, then the method in Example 2 can be used for the time synchronization.

Symbol Synchronization the aim of the symbol synchronization is to align symbols between the transmitter and the receiver so that the receiver can correctly make signal detection. Two examples are provided to explain the symbol synchronization in molecular communication systems.

Example 4 (ShahMohammadian et al. 2013) : In this example, MoSK is adopted, meaning that different molecules are used to represent different symbols. Because different molecules have different diffusion coefficient, the peak concentration time, which is defined as the propagation, will be different. It will be better to estimate this peak concentration time correctly and then take sample at the peak concentration to make decision for the signal detection. The peak concentration has the largest signal-to-noise ratio (SNR) among all possible samples. The estimation of the peak concentration time based on the received signal is considered as the process of the symbol synchronization.



Synchronization of Nanomachines, Fig. 2 Message exchange mechanism for diffusion-based clock synchronization. The transmitter sends synchronization molecules

with the period of T . For the i th round, the transmitter records the molecule releasing time instant $T_{tx,i}$, and the receiver records the receiving time instant $T_{rx,i}$

Example 5 (Jamali et al. 2017) : In this example, more practical challenges are considered. In the practical conditions, the transmitter may not be equipped with an internal clock and may not be able to emit molecules with a fixed release frequency. The released time, which is considered as the start of the symbol interval, varies according to the availability of food and energy for molecule generation, the process for molecule production, etc. How to estimate the start of a symbol interval is the issue of the symbol synchronization, which is needed for reliable detection (Jamali et al. 2017).

For both of the above two examples, ML method can be used for the symbol synchronization. A general description of the ML method for the symbol synchronization is as follows. It is assumed that $s(t)$ is the transmitted signal, $z(t)$ is the received signal, and ζ is the unknown parameter which needs to be estimated for the symbol synchronization. ζ can be either the peak concentration time (i.e., propagation delay) in Example 4 or the start time instant of a symbol interval Example 5. The metric $\Lambda_{ML}(\zeta)$ is the probability function of the received signal given the unknown, as

$$\Lambda_{ML}(\zeta) = p(z(t); \zeta). \quad (8)$$

An estimate of the unknown ζ_{est} can be expressed as

$$\zeta_{\text{est}} = \arg \max_{\zeta} \Lambda_{ML}(\zeta). \quad (9)$$

Key Applications

Different types of synchronizations of nanomachines have different application scenarios. The synchronization of oscillation and time/clock synchronization can be involved in the scenarios of the coordination of the nanomachines (i.e., nanorobots), where the nanomachines need to coordinate their behaviors such as simultaneous drug releasing. The time/clock synchronization is necessary and essential in the IoNT scenarios.

The symbol synchronization is the basis for the data transmission/communication between nanomachines.

Future Directions

There are still some challenges for the synchronization of nanomachines if more complicated or more practical scenario is taken into account. Two examples are listed below:

- The synchronization issue for the mobile molecular communication system is a challenge and has not been investigated. In that system, due to the motion of the nanomachines, the propagation delay becomes a random variable, which makes the synchronization difficult.
- A blood vessel channel is a realistic channel which is one of the potential scenarios for the molecular communication. The feature of the blood velocity makes the channel a time-varying channel, i.e., the CIR is dependent on the molecule releasing time. In this case, the realization of the synchronization becomes a challenge.

References

- Abadal S, Akyildiz IF (2011) Automata modeling of quorum sensing for nanocommunication networks. *Nano Commun Netw* 2(1):74–83
- Jamali V, Ahmadzadeh A, Schober R (2017) Symbol synchronization for diffusion-based molecular communications. *IEEE Trans NanoBiosci* 16: 873–887
- Kim NR, Chae CB (2013) Novel modulation techniques using isomers as messenger molecules for nano communication networks via diffusion. *IEEE J Sel Areas Commun* 31(12):847–856
- Lin L, Li F, Ma M, Yan H (2016a) Synchronization of bio-nanomachines based on molecular diffusion. *IEEE Sensors J* 16(19):7267–7277
- Lin L, Yang C, Ma M, Ma S, Yan H (2016b) A clock synchronization method for molecular nanomachines in bionanosensor networks. *IEEE Sensors J* 16(19): 7194–7203
- Lo C, Liang YJ, Chen KC (2014) A phase locked loop for molecular communications and computations. *IEEE J Sel Areas Commun* 32(12):2381–2391

- Luo Z, Lin L, Guo W, Wang S, Liu F, Yan H (2018) One symbol blind synchronization in simo molecular communication systems. *IEEE Wirel Commun Lett* 7(4):530–533
- Moore MJ, Nakano T (2013) Oscillation and synchronization of molecular machines by the diffusion of inhibitory molecules. *IEEE Trans Nanotechnol* 12(4):601–608
- Nakano T, Suda T, Okaie Y, Moore MJ, Vasilakos AV (2014) Molecular communication among biological nanomachines: a layered architecture and research issues. *IEEE Trans Nanobiosci* 13(3):169–197
- ShahMohammadian H, Messier GG, Magierowski S (2013) Blind synchronization in diffusion-based molecular communication channels. *IEEE Commun Lett* 17(11):2156–2159
- Srinivas K, Eckford AW, Adve RS (2012) Molecular communication in fluid media: the additive inverse gaussian noise channel. *IEEE Trans Inf Theory* 58(7):4678–4692

T

Tacit Collusion

- [Collusion](#)

Targeted Drug Delivery

- [Drug Delivery via Nanomachines](#)

Task Mapping

- [Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks](#)

TDMA Link Scheduling

- [Interference-Aware Scheduling in Wireless Networks](#)

Terahertz Channel Modeling

- [Channel Modeling of Microscale Terahertz Communication](#)

Terahertz-Band Nano-communications

- [Nanoscale Terahertz Communications](#)

The Draft IEEE 802.11i Standard

- [Wi-Fi Protected Access \(WPA\)](#)

The Primary Synchronization Sequences in LTE

Chao-Yu Chen
Department of Engineering Science, National
Cheng Kung University, Tainan, Taiwan

Synonyms

[Primary synchronization signal sequences](#)

Definitions

The primary synchronization signals in LTE are used to help user equipments achieve the time and frequency synchronization.

Historical Background

Cell search is an important procedure to establish the connection between enhanced NodeB (eNB) and user equipment (UE) before data transmission and reception in long-term evolution (LTE) standard specified by 3GPP (2017). Initial synchronization and cell identity (ID) identification can be achieved by the use of synchronization signals transmitted by eNBs. LTE provides the primary synchronization signal (PSS) and the secondary synchronization signal (SSS) to help UEs perform the timing and frequency synchronization as well as the cell ID detection. There are a total of 504 physical layer cell IDs in LTE networks (3GPP 2017). They are divided into 168 group IDs, denoted by $N_{ID}^{(1)}$, and each group contains 3 sector IDs, denoted by $N_{ID}^{(2)}$. The cell ID N_{ID}^{cell} is then given by

$$N_{ID}^{cell} = 3N_{ID}^{(1)} + N_{ID}^{(2)} \quad (1)$$

where $N_{ID}^{(1)}$ ranges from 0 to 167 and $N_{ID}^{(2)}$ ranges from 0 to 2. Three different Zadoff-Chu (ZC) sequences are used in PSSs to represent three different sector IDs, while Gold sequences are employed for SSSs.

Key Applications

The primary synchronization sequences can be used for time synchronization and frequency offset estimation and cell ID detection in the cell search procedure.

Zadoff-Chu Sequences

Zadoff-Chu sequences are named after Zadoff (Frank and Zadoff 1962) and Chu (1972), which are also called Frank-Zadoff-Chu (FZC) sequences. The ZC sequences have been widely used in LTE including PSSs, random access preambles, uplink reference signals, and physical uplink control channels (PUCCHs). Since ZC sequences have constant amplitudes and ideal

periodic autocorrelations, they are also referred to as constant amplitude zero autocorrelation (CAZAC) sequences. The properties of ZC sequences will be given later. A ZC sequence of length N_{ZC} can be defined as (Chu 1972).

$$x_u[n] = \begin{cases} e^{-j\frac{\pi un^2}{N_{ZC}}}, & \text{even } N_{ZC}; \\ e^{-j\frac{\pi un(n+1)}{N_{ZC}}}, & \text{odd } N_{ZC}, \end{cases} \quad (2)$$

for $n = 0, 1, \dots, N_{ZC} - 1$, where u is called the root index and is relatively prime to N_{ZC} , that is, $\gcd(u, N_{ZC}) = 1$. Some major properties of ZC sequences are given as follows:

Property 1

For any sequence $s = (s[0], s[1], \dots, s[L-1])$ of length L , the periodic autocorrelation at displacement p is defined as

$$\rho(s; p) = \sum_{k=0}^{L-1} s[(k+p) \bmod L] s^*[k], \quad (3)$$

for $p = 0, 1, \dots, L-1$

where $(k+p) \bmod L$ denotes the modulo- L addition of k and p and $(\cdot)^*$ denotes the complex conjugate.

For an ZC sequence $x_u = (x_u[0], x_u[1], \dots, x_u[N_{ZC}-1])$ of length N_{ZC} , it has ideal periodic autocorrelation (Frank and Zadoff 1962; Chu 1972), that is,

$$\rho(s; p) = \begin{cases} N_{ZC}, & p = 0; \\ 0, & p = 1, 2, \dots, N_{ZC} - 1. \end{cases} \quad (4)$$

In addition, the amplitude $|x_u[n]| = 1$ for all n , and hence the ZC sequence has the CAZAC property.

Property 2

For odd-length N_{ZC} , two ZC sequences with root indices u and $N_{ZC}-u$ have the complex conjugate property (Popovic and Berggren 2008):

$$x_{N_{\text{ZC}}-u}[n] = x_u^*[n], \quad \text{for } n = 0, 1, \dots, N_{\text{ZC}} - 1. \quad (5)$$

Therefore, the ZC sequence with root index $N_{\text{ZC}} - u$ can be easily obtained from $x_u[n]$.

Property 3

A ZC sequence of odd length is centrally symmetric (Popovic and Berggren 2008):

$$x_u[N_{\text{ZC}} - 1 - n] = x_u[n], \quad \text{for } n = 0, 1, \dots, (N_{\text{ZC}} - 1)/2. \quad (6)$$

Property 4

For two odd-length ZC sequences with different root indices u_1 and u_2 , they have the theoretical minimum cross-correlations (Sarwate 1979) $\sqrt{N_{\text{ZC}}}$ if $|u_1 - u_2|$ is relatively prime to N_{ZC} . That is, the absolute value of the cross-correlation function of $x_{u_1}[n]$ and $x_{u_2}[n]$ at shift p is obtained by (Popovic 1992)

$$\begin{aligned} & |\rho(\mathbf{x}_{u_1}, \mathbf{x}_{u_2}; p)| \\ &= \left| \sum_{k=0}^{N_{\text{ZC}}-1} x_{u_1}[(k+p) \bmod N_{\text{ZC}}] x_{u_2}^*[k] \right| \\ &= \sqrt{N_{\text{ZC}}}, \quad \text{for } p = 0, 1, \dots, N_{\text{ZC}} - 1. \end{aligned}$$

Considering a special case with a prime number N_{ZC} , any two distinct sequences of the set of $N_{\text{ZC}} - 1$ ZC sequences with root indices $u = 1, 2, \dots, N_{\text{ZC}} - 1$ have low and constant cross-correlations.

Property 5

If the sequence length N_{ZC} is a prime, there is a remarkable discrete Fourier transform (DFT) property of the ZC sequence. For a ZC sequence $x_u[n]$ with root index u , the N_{ZC} -DFT of $x_u[n]$ denoted by $X_u[n]$ can be expressed as a weighted, conjugated, and time-scaled version of $x_u[n]$ (Beyme and Leung 2009):

$$X_u[n] = x_u^*[u^{-1}n]X_u[0] \quad (7)$$

where u^{-1} is the modular multiplicative inverse of u , that is, $u^{-1}u \equiv 1 \pmod{N_{\text{ZC}}}$. Therefore, DFT operation is not required to obtain the DFT of a ZC sequence. It can be seen that $X_u[n]$ also has constant amplitudes.

Primary Synchronization Signals

The PSSs use three 63-length ZC sequences with root indices 25, 29, and 34 to represent the sector IDs $N_{\text{ID}}^{(2)} = 0, 1$, and 2, respectively (3GPP 2017). Since the ZC sequences are employed in the frequency domain with center element nulled to avoid the effect of DC offset, the primary synchronization sequence is generated as below:

$$d_u[n] = \begin{cases} e^{-j\frac{\pi u n(n+1)}{63}}, & \text{for } n=0, 1, \dots, 30; \\ e^{-j\frac{\pi u(n+1)(n+2)}{63}}, & \text{for } n=31, 32, \dots, 61. \end{cases} \quad (8)$$

And the sequence is mapped to the center 62 subcarriers skipping the DC subcarrier. That is, the k th subcarrier $a[k]$ is assigned by

$$a[k] = d_u[n], \quad \text{for } n = 0, 1, \dots, 61 \quad (9)$$

with

$$k = n - 31 + \left\lfloor \frac{N_{\text{RB}}^{\text{DL}} N_{\text{SC}}^{\text{RB}}}{2} \right\rfloor \quad (10)$$

where $N_{\text{RB}}^{\text{DL}}$ denotes the number of resource blocks (RBs) for downlink (DL) transmission and $N_{\text{SC}}^{\text{RB}}$ denotes the number of subcarriers in one RB. Six center RBs are reserved for PSSs on specified OFDM symbols, so there are 72 reserved subcarriers, while the sequence length of $d_u[n]$ is only 62. One major reason to use 63-length ZC sequence is that 64-point fast Fourier transform (FFT) with lower complexity can be adopted. Moreover, the sequence length is taken as an odd prime number, so the primary synchronization sequences can possess Property 2, Property 3, Property 4, and Property 5 of ZC sequences. For Property 2, since the root indices

29 and 34 satisfy $34 = 63 - 29$, the equation $d_{34}[n] = d_{29}^*[n]$ can be utilized in sector ID detection algorithm. Property 4 enables UEs to detect the correct sector ID by finding the maximum correlations due to the low cross-correlations between different ZC sequences. Although Property 5 tells us that the inverse discrete Fourier transform (IDFT) of a ZC sequence has constant amplitudes, the IDFT size should be equal to the length of the sequence. Since 64-FFT with low complexity is preferred and the center element of the ZC sequence is nulled, the time-domain signal of a PSS by using 64-FFT does not have constant amplitudes anymore. Note that although the peak-to-average power ratio (PAPR) of a PSS is not 0dB, the PAPRs are still low (Tsai et al. 2007).

According to the good properties of ZC sequences, they have also been employed as synchronization sequences in narrowband Internet of Things (NB-IoT) (3GPP 2017). The primary and secondary synchronization signals use ZC sequences of lengths 11 and 131, respectively, in NB-IoT.

Cross-References

- [Cellular Networks: An evolution from 1G to 4G](#)
- [Key Technologies in 4G/LTE Network](#)
- [Narrowband Internet of Things](#)

References

- 3GPP (2017) Evolved universal terrestrial radio access (E-UTRA) – physical channels and modulation. 3GPP TS 36.211 V14.5
- Beyme S, Leung C (2009) Efficient computation of DFT of Zadoff-Chu sequences. *Electron Lett* 45(9):461–462
- Chu DC (1972) Polyphase codes with good periodic correlation properties. *IEEE Trans Inf Theory* IT-18(4):531–532
- Frank RL, Zadoff SA (1962) Phase shift pulse codes with good periodic correlation properties. *IEEE Trans Inf Theory* IT-18(6):381–382
- Popovic BM (1992) Generalized chirp-like polyphase sequences with optimum correlation properties. *IEEE Trans Inf Theory* 38(4):1406–1409

- Popovic BM, Berggren F (2008) Primary synchronization signal in E-UTRA. In: *Proceedings of IEEE international symposium on spread spectrum technical and application*, Bologna, pp 426–430
- Sarwate DV (1979) Bounds on crosscorrelation and autocorrelation of sequence. *IEEE Trans Inf Theory* IT-25(6):720–724
- Tsai Y, Zhang G, Grieco D, Ozluturk F, Wang X (2007) Cell search in 3GPP long term evolution systems. *IEEE Veh Technol Mag* 2(2):23–29

Thermal Excitation of Molecules

- [Brownian Motion](#)

Threat, Attack

- [Wireless Sensor Network Security](#)

3D MU-Massive MIMO

- [Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System](#)

3GPP Release 15

- [5G Wireless](#)

3V

- [Ubiquitous Big Data](#)

Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System

Saeid Aghaeinezhadfirouzja and Bin Xia
Department of Electronics Engineering,
Shanghai Jiao Tong University, Shanghai,
People's Republic of China

Synonyms

3D MU-massive MIMO; Channel delay spread;
Power delay profile; Received spatial correlation

Definitions

This entry presents a practical three-dimensional (3D) outdoor-to-indoor channel model for multiuser massive multi-input multi-output (MU-massive MIMO). The model shows that the reflected clusters after passing through the boundary of the physical objectives from outdoor to indoor will be divided into several Ray paths, which results in different elevation/azimuth angles and phase excitation in all directions. The received spatial correlation is investigated based on the proposed 3D MU-massive MIMO. In addition, channel delay spread and multipath power delay profile (PDP) properties are studied based on the measured data. Finally, the result convinces that the proposed model is able to capture specific characteristic of outdoor-to-indoor 3D MU-massive MIMO.

Introduction

Recently, massive multi-input multi-output (MIMO) has attracted substantial research attentions as an important technique to address the demand for high data rate transmission. However, an accurate radio propagation model will be developed along the new technique.

The authors in 3GPP (July, 2015) provided an important example of the three-dimensional (3D) indoor hotspot channel model. This entry summarized the key results of indoor communications in addition to the outdoor contribution the 3rd Generation Partnership Project (3GPP) in Haneda et al. (2016b). In Weng et al. (2014), classified MIMO channels with a smaller number of antennas are investigated. Several experiments have been made to study the spatial correlation and the capacity indoor and outdoor-to-indoor environment in Kafle et al. (2008) and Kyritsi et al. (2003). In Kafle et al. (2008), for an instance, MIMO was designed for broadband MIMO channel sounding, capacity measurements, and characterizing the directional multipath of the radio propagation channel. In Kyritsi et al. (2003), the spatial correlation has been studied for in MIMO antenna system. In addition, the authors investigated an indoor 5G channel models for office and shopping mall environments in Haneda et al. (2016a). However, the impact of the elevation angles and the antenna element spacing for indoor 3D channel modeling was not considered in detail. In Yuan et al. (2015) and Renaudin et al. (2010), the work was based on the 3D theoretical channel model. The authors assumed an infinite number of effective scatters with resulting infinite complexity that can hardly be put into practical use. Furthermore, a novel 3D antenna polarized channel model was presented in Shafi et al. (2006). It investigated the impact of elevation angles for a MIMO system with a small number of antennas. Moreover, many researchers such as Dao et al. (2011) and Xu et al. (2004) investigated antenna configuration and polarization. A closed-form expression of the spatial correlation as a function of the physical parameters representing both characteristics of 3D MIMO was derived in Dao et al. (2011). Xu et al. (2004) presented a generalized MIMO channel modeling technique combining the spatial correlation and wave position in the direction of antenna arrays to achieve both accuracy and efficiency.

However, according to measurement observation in Payami and Tufvesson (2012), Gao et al. (2012), and Aghaeinezhadfirouzja et al. (2016, 2017), the abovementioned channel modeling

is not sufficient to accurately capture certain characteristics of 3D massive MIMO channels. Generally speaking, the previous researchers considered the channel modeling with a small number of antennas, indoor model with the impact on the horizontal or elevation angles by theoretical properties, not suitable to be directly applied to modeling 3D outdoor-to-indoor massive MIMO channels. In outdoor-to-indoor 3D massive MIMO environment, the clusters arrive from open to the close area passing through different scatters which will divided them into different Ray paths. Therefore, the Ray paths may obtain different elevation/azimuth angles based on their time delays and locations, where their angular of arrivals and departures need to be readdressed while generating the channel coefficients. Moreover, channels can vary with antenna configuration depending on the number of antenna elements used in massive MIMO system. It will then need properly model elevation angle of departures (EAoDs) and elevation angle of arrivals (EAoAs) as well as azimuth angle of departures (AAoDs) and azimuth angle of arrivals (AAoAs) of multipath components from the Ray paths that will help to contribute the spatial correlation among different antenna elements to the receivers (RXs). Therefore, to investigate the received spatial correlation, we considered antenna element space (AES) to recognize the antenna element location and the distance between two individual antenna elements for the correlation purposes.

In summary, we are introducing 3D outdoor-to-indoor channel models based on extensive measurements as well as preliminary result on channel parameters that are capable of capturing main properties of the 3D MU-massive MIMO.

The major contributions of this entry are the following:

1. In outdoor-to-indoor environments, the clusters arrive from open to close area passing through physical objectives on the outdoor-to-indoor boundaries, which will divide the clusters into several Ray paths and may provide new elevation and azimuth angles

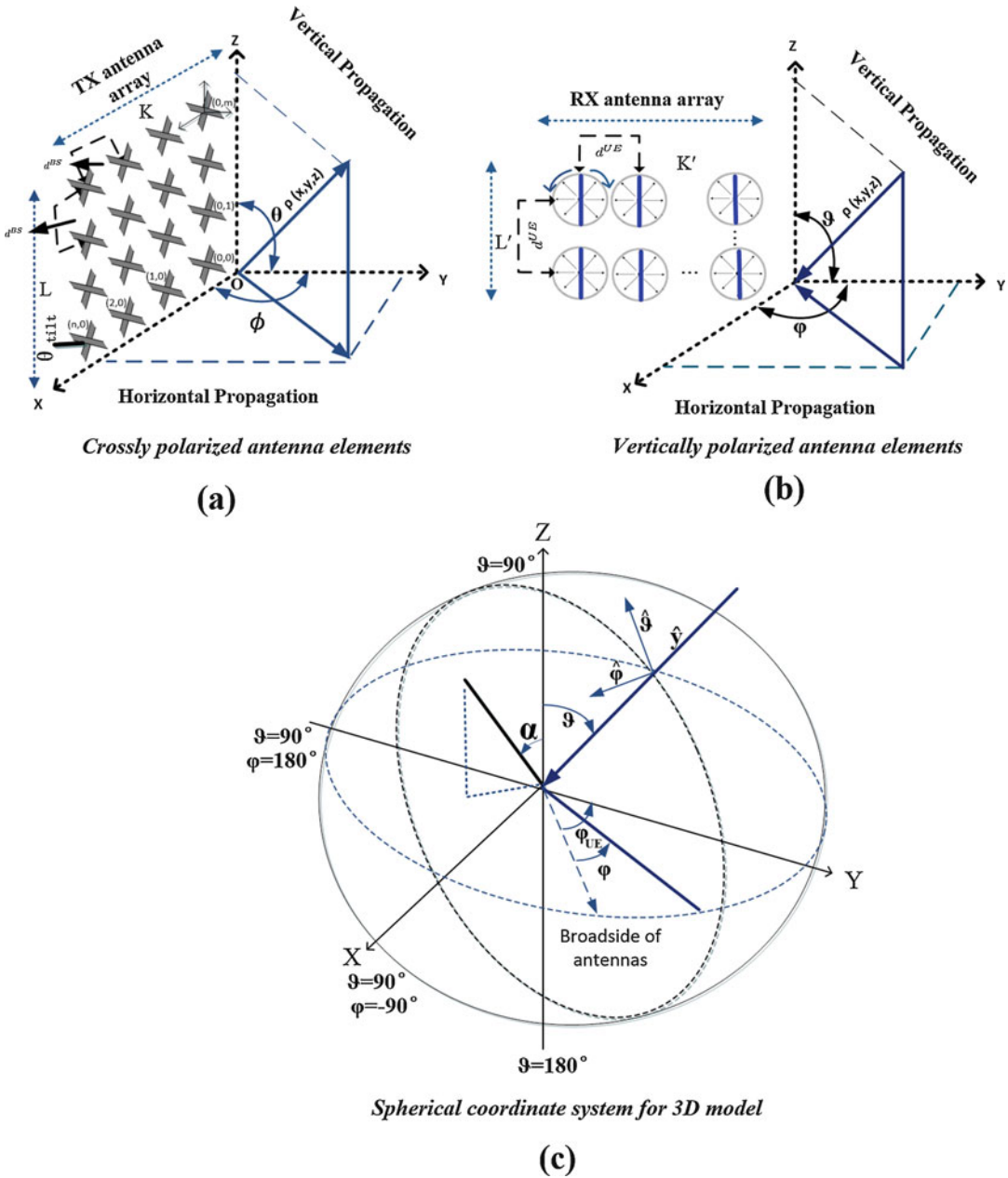
toward different directions. This entry first proposes a 3D outdoor model and then based on the outdoor model; 3D indoor model has been investigated. And then, we investigate the characterization of the reflecting clusters which will be generated/regenerated to several multipath signals in 3D. So far, few works have been done on the generating/regenerating of the clusters to multipath.

2. An interesting interplay between vertical antenna pattern and elevation spread is observed with AES model that requires capturing the characteristics of elevation/azimuth angles for both vertical and horizontal directions of the antenna arrays. The goal of this part is also to provide different phase directions on received spatial correlation.
3. Statistical properties of the proposed model such as power delay profile (PDP) and delay spread (DS), including multipath richness and channel capacity/correlation of the 3D nonstationary channel model, were investigated. The proposed model is valid for outdoor-to-indoor communication in 3D MU-massive MIMO scenarios with the measured channel data.

Antenna Pattern

In this section, downlink 3D MU-massive MIMO system and polarization mixing were considered for outdoor-to-indoor situation. Massive MIMO antenna with pairs of cross-dipole elements with ± 45 degrees of polarization slant angle was installed at the transmitter (TX). Consider the polarized massive MIMO shown in Fig. 1, where (a) cross polarized (X-Pol) dipole elements at the TX, (b) is with vertical polarized (V-Pol) omnidirectional antenna elements at the RX, and (c) is the beam pattern coordinate system for 3D model.

Let us denote the elevation angle as θ and the azimuth angle as ϕ at the TX. Similarly ϑ is the elevation angle and φ is azimuth angle at the RX. The transmit massive MIMO antenna arrays are fixed at the $x - z$ plane both for transmitter and receiver. The minimal distances between antenna elements are d^{BS} and d^{UE} both for the TX and



Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System, Fig. 1 Two antenna patterns: 3D channel model of antenna elements in the hor-

izontal and vertical directions is indexed by (l, k) and (l', k') for (a) X-Pol and (b) V-Pol polarizations, respectively, and (c) beam pattern coordinate system for 3D model

the RX antennas, respectively. Denote l and k as the horizontal and the vertical indexes for $(l, k)_{th}$ of the transmit antenna elements, where $l = \{0, \dots, L-1\}$ and $k = \{0, \dots, K-1\}$, i.e., there are $L \times K$ pairs of transmit antenna elements.

Similarly, denote l' and k' as the horizontal and the vertical positions of the antenna elements, where $l' = \{0, \dots, L'-1\}$ and $k' = \{0, \dots, K'-1\}$, i.e., there are $L' \times K'$ pairs of antenna elements at the receiver. Let the location of the

$(0, 0)_{th}$ TX antenna element be the origin point O ; the location of the $(0, 0)_{th}$ receive antenna element is $(x_{(0,0)}, y_{(0,0)}, z_{(0,0)})$. Then we can determine the location and the distance of the other antenna elements to the $(0, 0)_{th}$ antenna element of the antenna array as follows.

Transmit Antenna Array

Then, the distance between the $(l, k)_{th}$ antenna element and the $(0, 0)_{th}$ antenna element at the transmitter can be computed as

$$d_{l,k}^{BS} = \sqrt{l^2 + k^2} \cdot d^{BS} \quad (1)$$

Moreover, vector $\mathbf{a}(\phi, \theta)_{l,k}$ is the array responses of the $(l, k)_{th}$ TX antenna element which is given by

$$\mathbf{a}(\phi, \theta)_{l,k} = e^{\left(\frac{2\pi}{\lambda} \hat{y} \cdot d_{l,k}^{BS}\right)} \quad (2)$$

where λ is the wavelength of the carrier frequency. $\hat{y} = [\sin \phi_{AOD} \cos \theta_{AOD}, \sin \phi_{AOD} \sin \theta_{AOD}, \cos \theta_{AOD}]$ is the outgoing wave direction.

Receive Antenna Array

Similarly, the distance between the $(l', k')_{th}$ antenna element and the $(0, 0)_{th}$ antenna element at the receiver can be computed as

$$d_{l',k'}^{UE} = \sqrt{l'^2 + k'^2} \cdot d^{UE} \quad (3)$$

The polarization vector at angle α from the z -axis in Fig. 1c has vertical and horizontal components of the antenna pattern, by expressing the response vector X in its ϑ and ϕ components which are proportional to Shafi et al. (2006)

$$\begin{aligned} X_{l',k'}^{UE} &= \begin{bmatrix} X_{l',k'}^{\vartheta} \\ X_{l',k'}^{\phi} \end{bmatrix} \\ &= \begin{bmatrix} \cos \alpha \sin \vartheta + \cos \alpha \cos \vartheta \cos \phi \\ \sin \alpha \cos \phi \end{bmatrix} e^{\left(\frac{2\pi}{\lambda} \hat{y}' \cdot d_{l',k'}^{UE}\right)} \end{aligned} \quad (4)$$

where X^{ϑ} and X^{ϕ} are the ϑ and ϕ polarized responses of the antenna at the direction of $\hat{y}' =$

$[\sin \phi \cos \vartheta, \sin \phi \sin \vartheta, \cos \vartheta]$ for the incoming wave in the response of the antenna. Therefore, the vector $\mathbf{a}(\varphi, \vartheta)_{l',k'}$ is the array response of the $(l', k')_{th}$ receive RX antenna element, and the $(0, 0)_{th}$ antenna element is given by

$$\begin{aligned} \mathbf{a}(\varphi, \vartheta)_{l',k'} &= \begin{bmatrix} \cos \alpha \sin \vartheta + \cos \alpha \cos \vartheta \cos \phi \\ \sin \alpha \cos \phi \end{bmatrix} \\ &e^{\left(\frac{2\pi}{\lambda} \hat{y}' \cdot d_{l',k'}^{UE}\right)} \end{aligned} \quad (5)$$

A Practical Model for Outdoor-to-Indoor Massive MIMO Systems

In this section, a 3D channel model for MU-massive MIMO is proposed, considering spherical wave front and cluster/Ray path evaluation in both time and array axes as shown in Fig. 2. The models are defined by two effective physical objectives, one around the TX and other one in the cube at the office/shopping mall around the RX. Prior to this, we assumed that the clusters reflect from different sides of the cube which are divided into several Ray paths that contribute different elevation/azimuth angles. In addition, we used a spherical model for dipole antennas for outdoor communication and a cube model for indoor communication to minimize clusters/Ray paths dispersing. The users are randomly located inside the cube. Some important characteristics of massive MIMO outdoor-to-indoor channel models are needed to be investigated, such as the distance between TX and RX, cluster appearance on the antenna array for outdoor situation, generating/regenerating of Ray path models for indoor situation, and Ray appearance on the RX antenna array.

Therefore, to generate the channel coefficient, three parts must be modeled when signals are traveling from the transmitter toward the receiver such as (a) toward the outdoor cluster, (b) toward the cube boundary (points), and (c) toward the receive antenna elements, respectively:

- Toward the outdoor cluster: Let us denote the $qt = \{1, \dots, Qt\}$, where Qt is equal to $\text{Cluster}_t = Qt(t' + \Delta t')$ as total clusters, in

which i'_{th} path from $(l, k)_{th}$ antenna element is added up with time $(t' + \Delta t')$. Prior to this, the first path from $(0, 0)_{th}$ antenna element with time t' is added up to the nearest cluster. And then, the rest of the paths from $(l, k)_{th}$ adjacent element are calculating with distance $d_{l,k}^{BS}$ which are added up with succeeding time instantaneously $(\Delta t')$ to the same cluster. The distance between the transmit antenna and the cube array is denoted by D .

- Toward the cube boundaries (points): q_{th} cluster is observable to the $q_c = \{1, \dots, Q_c\}$ cluster, where Q_c is equal to $\text{Cluster}_{cube} = Q_c(t')$ as total clusters at the cube boundary based on distance D between the transmit antenna array and the cube boundary. In addition, there are s_{th} points, where $s = \{1, \dots, S\}$ are randomly located on the cube boundaries in which the cluster is passing through them toward the receive antenna elements.
- Toward the receive antenna elements: q_{th} cluster after passing through s_{th} point will be divided into several Ray paths. As an example, cluster q_{c1} is divided to the q_{th} Ray path, where $q = \{1, \dots, Q\}$. Therefore, denote $\text{Ray}_r = Q_r(t' + \Delta t')$ as total Ray paths which are observed at the l', k' antenna elements, where first Ray path is arriving with time t' toward $(0, 0)_{th}$ antenna element and then the rest of the Ray paths are arriving with succeeding time instantaneously $(\Delta t')$ toward $(l', k')_{th}$ antenna element that is calculating with distance $d_{l',k'}^{UE}$. The distance between the receive antenna array and the cube array is denoted by D_c . Therefore, the total distances between the transmit antenna array and the receive antenna array can be calculated by $D' = (D + D_c, 0, 0)$.

A 3D Massive MIMO Channel for outdoor Scenario

To perform the evolution of outdoor channel coefficient, the clusters in the set of $(Q_t(t' + \Delta t') \cap Q_c(t'))$ are generated based

on the outdoor cluster evolution for outdoor communication. Based on the outdoor model, the multipath gain between the transmit antenna and the cube boundary will be denoted as (CI_n^{out}) which is the total number of clusters, where $n = 1, \dots, N$ cluster for outdoor communication that can be expressed as

$$CI_n^{\text{out}} = \text{card} \left(\bigcup_{l,k=1}^{L,K} \bigcup_{s=1}^S \left(Q_t(t' + \Delta t') \cap (Q_c(t')) \right) \right) \quad (6)$$

where the operator $\text{card}(\cdot)$ denotes the cardinality of a set (Wu et al. (2014, 2015)). The operator $\text{card } CI_n^{\text{out}}$, each cluster, says $\text{Cluster}_n (n = 1, \dots, CI_n^{\text{out}})$ is for outdoor models.

Based on the above analysis and the summary of key parameter definitions in Table 1, the massive MIMO channel matrix can be expressed as an $L, K \times S$ complex matrix $H(t', \Delta t', \tau) = [h_{l,k,s}(t' + \Delta t', \tau)]_{L,K \times S}$. The amplitude complex gains between the $(l, k)_{th}$ antenna elements at the L, K transmit antenna, s_{th} point at the cube boundary and the delay τ , can be represented as

$$H_{l,k,s}(r' + \Delta t', \tau) = \sum_{n=1}^{CI_n^{\text{out}}} H_{l,k,n,s}(r' + \Delta t') \delta(\tau - \tau_n) \quad (7)$$

The calculation of complex gains for outdoor model can be expressed as LOS and NLOS components.

1. NLOS, part (A): The $(l, k)_{th}$ transmit antenna element vectors $\mathbf{a}(\phi, \theta)_{l,k}$ obtained by Eq. (2) in Section “A 3D Massive MIMO Channel for outdoor Scenario” and the vector between n_{th} cluster via i'_{th} path can be presented as $D_{n,i',l,k}^{BS}(t' + \Delta t')$, and the vector between the n_{th} cluster and transmit antenna array is denoted by $D_n^{BS}(t')$ at the TX.

Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System, Table 1
Summary of key parameter definitions

Parameters	Values
$(\theta, \phi), (\vartheta, \varphi)$	Elevation and azimuth angles of the TX and RX arrays, respectively
$D_{n,i',l,k}^{BS}(t' + \Delta t'), D_{LOS,i',l,k}^{BS}(t' + \Delta t')$	Distance vector between n_{th} and LOS_{th} cluster and $(l, k)_{th}$ transmit antenna elements via i'_{th} path, respectively
$D_n^{BS}(t'), D_{LOS}^{BS}(t')$	Distance vector between n_{th} and LOS_{th} cluster and $(l, k)_{th}$ transmit antenna elements, respectively
$D_{n,s}^{cube}(t'), D_{LOS,s}^{cube}(t')$	Distance between s_{th} vector and n_{th}, LOS_{th} cluster at cube boundary, respectively
$D_{s,q,l',k'}^{UE}(t' + \Delta t'), D_{s,LOS,l',k'}^{UE}(t')$	Distance vector between q_{th} and LOS_{th} Ray path and $(l', k')_{th}$ receive antenna elements from s_{th} point, respectively
$D_{s,l',k'}^{UE}(t'), D_s^{UE}(t')$	Distance vector between q_{th} and LOS_{th} Ray path and $(l', k')_{th}$ receive antenna elements, respectively
$\mathbf{a}(\phi, \theta)_{l,k}, \mathbf{a}(\varphi, \vartheta)_{l',k'}$	Array response of the TX (X-Pol) and RX (V-Pol) polarization
$\rho_{l,k,l',k',n,s,q}, \rho_{l,k,l',k',s}$	Received spatial correlation function of the received signal both for NLOS and LOS components, respectively
$\Phi_q^{(v,v)}, \Phi_q^{(v,h)}, \Phi_q^{(h,v)}, \text{ and } \Phi_q^{(h,h)}$	Four random initial phases (vertical/vertical), (vertical/horizontal), (horizontal/vertical), and (horizontal/horizontal), respectively

$$\begin{aligned} [D_{n,i',l,k}^{BS}(t' + \Delta t')] &= D_n^{BS}(t') \\ \begin{bmatrix} \sin \phi_{n,i',l,k}^{AAoD}(t' + \Delta t'), \cos \theta_{n,i',l,k}^{EAoD}(t' + \Delta t') \\ \sin \phi_{n,i',l,k}^{AAoD}(t' + \Delta t'), \sin \theta_{n,i',l,k}^{EAoD}(t' + \Delta t') \\ \cos \theta_{n,i',l,k}^{EAoD}(t' + \Delta t') \end{bmatrix}^T &= \begin{bmatrix} \exp(j\Phi_n^{(v,v)}) & \sqrt{X^h} \exp(j\Phi_n^{(v,h)}) \\ \sqrt{X^v} \exp(j\Phi_n^{(h,v)}) & \exp(j\Phi_n^{(h,h)}) \end{bmatrix} \end{aligned} \quad (8)$$

Similarly, the s_{th} vector and n_{th} cluster in the direction of waves at the cube boundary can be presented as

$$[D_{n,s}^{cube}(t')] = \begin{bmatrix} \sin \phi_{n,s}(t'), \cos \theta_{n,s}(t') \\ \sin \phi_{n,s}(t'), \sin \theta_{n,s}(t') \\ \cos \theta_{n,s}(t') \end{bmatrix} + D \quad (9)$$

Then, the four random initial phases for n_{th} cluster are derived as

where $\sqrt{X^h}$ and $\sqrt{X^v}$ are inverse crosses polarized for X-Pol (vv/hv and hh/vh) transmission, respectively. $j = \sqrt{-1}$ and $\Phi_n^{(v,v)}, \Phi_n^{(v,h)}, \Phi_n^{(h,v)}, \Phi_n^{(h,h)}$ are the four random initial phases of different polarization combinations.

2. LOS, part (A): The $(l, k)_{th}$ transmit antenna element vectors $\mathbf{a}(\phi, \theta)_{l,k}$ and the vector between LOS_{th} cluster via i'_{th} path are denoted by $D_{LOS,i,l,k}^{BS}(t' + \Delta t')$, and the vector between the LOS_{th} cluster and the transmit antenna array is denoted by $D_{LOS}^{BS}(t')$ at the TX.

$$[D_{LOS,i',l,k}^{BS}(t' + \Delta t)] = D_{LOS}^{BS}(t') \begin{bmatrix} \sin \phi_{LOS,i',l,k}^{AAoD}(t' + \Delta t'), \cos \theta_{LOS,i',l,k}^{EAoD}(t' + \Delta t') \\ \sin \phi_{LOS,i',l,k}^{AAoD}(t' + \Delta t'), \sin \theta_{LOS,i',l,k}^{EAoD}(t' + \Delta t') \\ \cos \theta_{LOS,i',l,k}^{EAoD}(t' + \Delta t') \end{bmatrix}^T \quad (11)$$

Similarly, the distance s_{th} vector and LOS cluster vector at cube boundary in the direction of waves can be presented as

$$\begin{aligned} & [D_{LOS,s}^{\text{cube}}(t')] \\ &= \begin{bmatrix} \sin \phi_{LOS,s}(t'), \cos \theta_{LOS,s}(t') \\ \sin \phi_{LOS,s}^{AAoD}(t'), \sin \theta_{LOS,s}(t') \\ \cos \theta_{LOS,s}(t') \end{bmatrix} + D \end{aligned} \quad (12)$$

Then, the two random initial phases for outdoor $outLOS_{th}$ cluster is derived as

$$\begin{bmatrix} \exp(j\Phi_{outLOS}^{(v,v)}) & 0 \\ 0 & \exp(j\Phi_{outLOS}^{(h,h)}) \end{bmatrix} \quad (13)$$

where $\Phi_{outLOS}^{(v,v)}$ and $\Phi_{outLOS}^{(h,h)}$ are two random initial phases of different polarization combinations.

Next, let us denote Rician factor as G . Additionally, the power of the n_{th} cluster is defined as A_n . Therefore, a $\text{Clusters}_n \in (Qt(t' + \Delta t) \cap Qc(t')) \leq Cl_n^{\text{out}}$ is observed for outdoor model.

– if $\text{Clusters}_n \in \cup_{l,k=1}^{L,K} \cup_{s=1}^S (Qt(t' + \Delta t) \cap Qc(t'))$,

$$H_{l,k,s,n}(t' + \Delta t', \tau) =$$

$$\begin{aligned} & \left(\delta(n-1) \sqrt{\frac{G}{G+1}} [D_{LOS,i',l,k}^{BS}(t' + \Delta t')] \times \begin{bmatrix} \exp(j\Phi_{outLOS}^{(v,v)}) & 0 \\ 0 & \exp(j\Phi_{outLOS}^{(h,h)}) \end{bmatrix} \times [D_{LOS,s}^{\text{cube}}(t')] \right) \\ & \times \mathbf{a}(\phi, \theta)_{l,k} + \sqrt{\frac{A_n}{G+1}} \sum_{n=1}^N [D_{n,i',l,k}^{BS}(t' + \Delta t')] \times \begin{bmatrix} \exp(j\Phi_n^{(v,v)}) & \sqrt{X^h} \exp(j\Phi_n^{(v,h)}) \\ \sqrt{X^v} \exp(j\Phi_n^{(h,v)}) & \exp(j\Phi_n^{(h,h)}) \end{bmatrix} \\ & \times [D_{n,s}^{\text{cube}}(t') \times \mathbf{a}(\phi, \theta)_{l,k}] \end{aligned} \quad (14)$$

– If $\text{Clusters}_n \notin \cup_{l,k=1}^{L,K} \cup_{s=1}^S (Qt(t' + \Delta t') \cap Qc(t'))$,

$$Cl_q^{\text{in}} = \text{card} \left(\cup_{s=1}^S \cup_{l',k'=1}^{L',K'} (Qc(t') \cap Qr(t' + \Delta t)) \right) \quad (16)$$

$$H_{l,k,s,n}(t' + \Delta t', \tau) = 0 \quad (15)$$

and the total number of clusters/Ray paths in Eqs. (6) and (16) for outdoor-to-indoor communications is derived as

$$Cl_{n,q}^{\text{Total}} = (Cl_n^{\text{out}} \cup Cl_q^{\text{in}}) \quad (17)$$

A 3D Massive MIMO for Outdoor-to-Indoor Scenario

In addition, to perform the evolution of channel coefficient for indoor communication, the Ray paths in the set $(Qc(t') \cap Qr(t' + \Delta t'))$ are generated based on the outdoor cluster after passing via s_{th} point and will be divided into the several Ray paths. Based on the indoor model, the multipath gain between cube boundary and the receive antenna elements, which will be denoted as (Cl_q^{in}) , is the total number of Ray paths on indoor communication that can be expressed as

where the operator $\text{card } Cl_q^{\text{in}}$, each Ray path, say $\text{Ray}_q (q = 1, \dots, Cl_q^{\text{in}})$, is for indoor models. Similarly, the massive MIMO channel matrix can be expressed as an $S \times L', K'$ complex matrix $H(t', \Delta t', \tau) = [h_{l,k,q}(t' + \Delta t', \tau)]_{S \times L', K'}$. The amplitude complex gains between the s_{th} point at the cube boundary and $(l', k')_{th}$ antenna elements at the L', K' receive antenna elements and delay τ can be represented as

$$H_{l',k',q}(t' + \Delta t', \tau) = \sum_{q=1}^{Cl_q^{in}} H_{l',k',q}(t' + \Delta t') \delta((\tau - \tau_q)(t')) \quad (18)$$

The calculation of complex gains for indoor model can expressed as LOS and NLOS components.

$$\left[D_{s,q,l',k'}^{UE}(t' + \Delta t') \right] = D_{q,l',k'(t')}^{UE} \begin{bmatrix} \sin \varphi_{s,q,l',k'}^{AAoA}(t' + \Delta t'), \cos \vartheta_{s,q,l',k'}^{AAoA}(t' + \Delta t') \\ \sin \varphi_{s,q,l',k'}^{AAoA}(t' + \Delta t'), \sin \vartheta_{s,q,l',k'}^{AAoA}(t' + \Delta t') \\ \cos \vartheta_{s,q,l',k'}^{AAoA}(t' + \Delta t') \end{bmatrix} + D_c \quad (19)$$

Then, the four random initial phases for indoor q_{th} Ray path are derived as

$$\begin{bmatrix} \exp(j\Phi_q^{(v,v)}) & \sqrt{X^h} \exp(j\Phi_q^{(v,h)}) \\ \sqrt{X^v} \exp(j\Phi_q^{(h,v)}) & \exp(j\Phi_q^{(h,h)}) \end{bmatrix} \quad (20)$$

where $\sqrt{X^h}$ and $\sqrt{X^v}$ are inverse polarizations in the cube. $j = \sqrt{-1}$ and $\Phi_q^{(v,v)}, \Phi_q^{(v,h)}$,

$$\left[D_{s,LOS,l',k'}^{UE}(t') \right] = D_s^{UE}(t') \begin{bmatrix} \sin \varphi_{s,LOS,l',k'}^{AAoA}(t' + \Delta t'), \cos \vartheta_{s,LOS,l',k'}^{AAoA}(t' + \Delta t') \\ \sin \varphi_{s,LOS,l',k'}^{AAoA}(t' + \Delta t'), \sin \vartheta_{s,LOS,l',k'}^{AAoA}(t' + \Delta t') \\ \cos \vartheta_{s,LOS,l',k'}^{AAoA}(t' + \Delta t') \end{bmatrix}^T + D_c \quad (21)$$

Then, the two random initial phases for indoor $inLOS_{th}$ Ray path are derived as

$$\begin{bmatrix} \exp(j\Phi_{inLOS}^{v,v}) & 0 \\ 0 & \exp(j\Phi_{inLOS}^{h,h}) \end{bmatrix} \quad (22)$$

1. NLOS, part (B): The $(l', k')_{th}$ receive antenna element vectors on polarization of $\mathbf{a}(\varphi, \vartheta)_{l',k'}$ obtained by Eq. (5) and the vector between q_{th} Ray path through s_{th} point can be denoted as $D_{s,q,l',k'}^{UE}(t' + \Delta t')$. Furthermore, the vector between the q_{th} Ray path and the receive antenna array is denoted by $D_{q,l',k'}^{UE}(t')$.

2. LOS, part (B): Similarly, the $(l', k')_{th}$ receive antenna element vectors for antenna polarization $\mathbf{a}(\varphi, \vartheta)_{l',k'}$ obtained by Eq. (5) and the vector between the LOS_{th} Ray path via s_{th} point can be denoted as $D_{s,LOS,l',k'}^{UE}(t' + \Delta t')$. Furthermore, the vector between the receive antenna array and s_{th} point is denoted by $D_s^{UE}(t')$.

where $\Phi_{inLOS}^{(v,v)}, \Phi_{inLOS}^{(h,h)}$ are random initial phases of different polarization combinations.

Therefore, a $\text{Ray}_q \in (Qc(t'), Qr(t' + \Delta t) \leq Cl_q^{in} \leq Cl_{n,q}$ is observed for indoor model.

If $\text{Ray}_q \in \cup_{s=1}^S \cup_{l',k'=1}^{L',K'} (Qc(t') \cap Qr(t' + \Delta t'))$,

$$H_{l',k',s,q}(t' + \Delta t', \tau) = \left(\begin{aligned} & \delta(q-1) \sqrt{\frac{G}{G+1}} \left[D_{s,LOS,l',k'}^{UE}(t') \right] \times \begin{bmatrix} \exp(j\Phi_{LOS}^{v,v}) & 0 \\ 0 & \exp(j\Phi_{inLOS}^{h,h}) \end{bmatrix} \times \left[D_{inLOS,s}^{\text{cube}}(t') \right] \\ & \times \mathbf{a}(\varphi, \vartheta)_{l',k'} + \sqrt{\frac{A_q}{G+1}} \sum_{q=1}^Q \left[D_{s,q,l',k'}^{UE}(t' + \Delta t') \right] \times \begin{bmatrix} \exp(j\Phi_q^{(v,v)}) & \sqrt{X^h} \exp(j\Phi_q^{(v,h)}) \\ \sqrt{X^v} \exp(j\Phi_q^{(h,v)}) & \exp(j\Phi_q^{(h,h)}) \end{bmatrix} \\ & \times \left[D_{n,s}^{\text{cube}}(t') \right] \times \mathbf{a}(\varphi, \vartheta)_{l',k'} \end{aligned} \right) \quad (23)$$

$$\begin{aligned}
& - \text{ If } \text{Ray}_q \notin \cup_{s=1}^S \cup_{l',k'=1}^{L',K'} (Qc(t') \\
& \cap Qr(t' + \Delta t')), \\
& H_{l',k',s,q}(t' + \Delta t', \tau) = 0
\end{aligned} \quad (24)$$

where the power of the q_{th} Ray path is defined as A_q . In the case of indoor communication, which is in the close area, one can assume that $G = \infty$, and for the case which is close to the window, a reasonable value for G is -4 dB Knudsen and Pedersen (2002).

Channel Properties

For the proposed 3D MU-massive MIMO, statistical properties will be derived in this section, i.e., received spatial correlation, multipath delay properties, and channel capacity. Here, we focused on the downlink channel model to quantify the effect of correlation and measure the performance gains realizable through potential elevation at the RX.

Outdoor-to-Indoor Multipath Delay Properties

In this subsection, multipath delay properties are proposed based on the 3D models. In this model, we assumed that there are two parts to investigate the delay properties, one for outdoor situation, from the TX toward the cube, and the other one is from the cube toward the users. We considered cluster and multipath time delays which are reflection with the delay of τ , each with the same power in indoor channel models. It is observed from indoor channel measurements that different clusters are divided into several multiple Ray paths and the arrival time clusters follow an individual Poisson process with the delay of each path spaced in an arbitrary manner. The arrival time of Ray in n_{th} cluster, denoted by T_n , is modeled by a Poisson process which is derived in Cho et al. (2010)

$$\begin{aligned}
f_{T_n}(T_n | T_{(n-1),n}) &= \Lambda \exp[-\Lambda(T_n - T_{(n-1)})], \\
n &= 1, 2, \dots
\end{aligned} \quad (25)$$

$$\begin{aligned}
f_{\tau_{q,n}}(\tau_{q,n} | \tau_{(q-1),n}) &= b \exp \\
&[-B(\tau_{q,n} - \tau_{(q-1),n})], \quad (26) \\
q &= 1, 2, \dots
\end{aligned}$$

where Λ and b are average arrival rate of cluster and Ray in each cluster, respectively. The $\tau_{q,n}$ denotes the arrival time of n_{th} cluster to the q_{th} Ray path. The arrival time of the n_{th} cluster, τ_n , is defined as the arrival time of the n_{th} cluster T_n .

Received Spatial Correlation and Capacity

In practice, the channels between the different antenna elements are often correlated, and therefore, the potential multiple antenna gains may not always be attainable (Wang and Murch (2006)). We focused and analyzed the spatial correlation of the RX, including the effect of antenna polarization based on the practical multipath indoor wireless communication environment. We will then investigate and minimize such correlation at the RX by considering the AES to recognize the antenna element and its adjacent antenna elements. Because of this, we developed received spatial correlation based in Cho et al. (2010). $\alpha_{n,q}$ and $\beta_{n,q}$ denote the amplitude and phase of q_{th} Ray path for v polarized antennas. Here, the received spatial correlation between two antenna elements via q_{th} path and the n_{th} cluster can be represented for both

$$H1_{l,k,l',k',n,s,q}(t' + \Delta t', \tau) = \alpha_q e^{j\beta_q} \sqrt{P}(\varphi, \vartheta) \quad (27)$$

$$\begin{aligned}
H2_{l,k,l',k',n,s,q}(t' + \Delta t', \tau) \\
= \alpha_q (\mathbf{a}(\varphi, \vartheta)_{l',k'} + \beta_q) \sqrt{P}(\varphi, \vartheta)
\end{aligned} \quad (28)$$

where $\sqrt{P}(\varphi, \vartheta)$ denotes the power angular spread.

However, the time difference between $H1$ and $H2$ as shown in Eqs. (27) and (28) and the following received spatial correlation function for the SULA with time instant t where all sub-Rays have equal power which can be expressed as

$$\rho_{l,k,l',k',n,s,q}(\lambda; t) = E \left[\frac{H1_{l,k,l',k',n,s,q}(t) H2_{l,k,l',k',n,s,q}^*(t)}{H1_{l,k,l',k',n,s,q}(t) H2_{l,k,l',k',n,s,q}^*(t)} \right] \quad (29)$$

As the result, the received spatial correlation can be rewritten as the sum of received spatial correlation of the *LOS* component $\rho_{l,k,l',k'}^{LOS}(\lambda; t)$ and the received spatial correlation of the *NLOS* components $\rho_{l,k,l',k',s}^{NLOS}(\lambda; t)$

$$\begin{aligned} \rho_{l,k,l',k',n,s,q}(\lambda; t) &= \rho_{l,k,l',k',s}^{LOS}(\lambda; r) \\ &+ \rho_{l,k,l',k',n,s,q}^{NLOS}(\lambda; t) \end{aligned} \quad (30)$$

For the received spatial correlation of the *LOS* component,

$$\rho_{l,k,l',k',s}^{LOS}(\lambda; r) = \frac{G}{G+1} \times e^{(j \frac{2\pi}{\lambda} \cdot d_{l',k'}^{UE} - \frac{2\pi}{\lambda} \cdot d_{l',k'}^{UE})} \quad (31)$$

In addition, the received spatial correlation among channel coefficient is fully characterized by the correlation matrix. Under the assumption, the ergodic channel capacity C for different polarization antennas, with their signal-to-noise ratio (SNR), is computed by using the following expression:

$$SNR = \frac{H^T H}{\sigma_n^2} \quad (32)$$

where σ_n^2 is the noise variance and H is the normalized channel matrixes, where channel capacity can be computed as

$$C = \log_2 [\det(I + SNR)] \text{ bps/Hz} \quad (33)$$

where $\det(\cdot)$ denotes the determinant and I is the identity matrix.

An Experimental Result and Discussion

This section provides the result of the real scenario of the proposed model from the measured data. An investigation has been made to study full characterization through a practical on 3D MU-massive MIMO systems, at Shanghai Jiao Tong University (SJTU) campus network, as shown in Fig. 3. UEs, each has omnidirectional elements, were utilized at the RXs. The location of UE1 is near the window, while UE2 and UE3 are in the same floor but both far from the window. The BS height about 35 m from the ground and the receiver antenna's heights in all UEs are 30 m in the office. The distribution of azimuth and elevation angles at the TX and RX has been studied, where azimuth and elevation angles are reflected in different directions from outdoor-to-indoor communications.

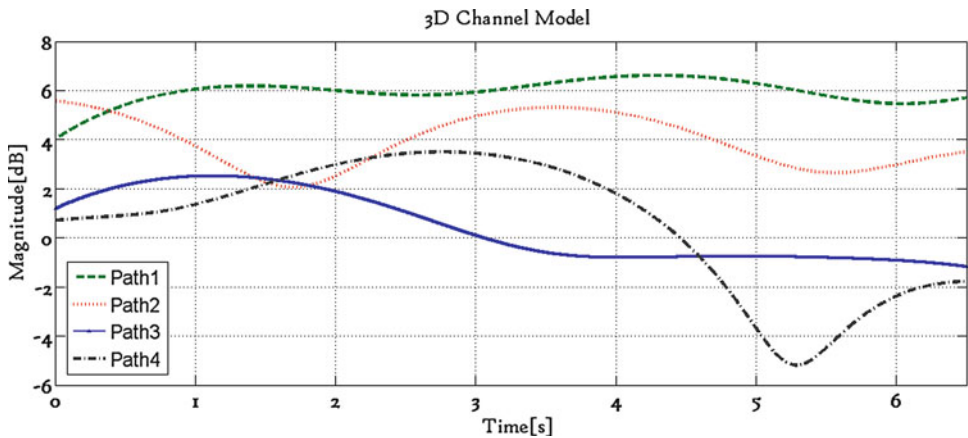
To perform our proposed model, delay properties and channel capacity/correlation of clusters/Ray paths were investigated according to the measured data and proposed model. The result was based on EAoA (ϑ), AAoA (φ), n_{th} cluster, (q_{th}) Ray path, element spacing, and the antenna polarization. Consequently, as shown in the following figures, we also investigated the performance of channel modeling with the effect of dual antenna polarization antenna. This observation also was from the dynamic property of clusters on the array axis which may observe different Ray paths with different (ϑ , φ) to the Rx.

Due to the presence of multiple electromagnetic paths at the RX side, more than one pulse is received. Figure 4 shows the time-domain channel coefficients between TX and UE1 which are generated by the time-domain models for proposed 3D MU-massive MIMO channel. As shown in the figure, path 1 is the *LOS* Ray path and the other are reflected paths. We proposed that a reflected cluster from physical objective be divided into several Ray paths which are observable to the user where he/she is located inside the cube. It is illustrated that the channel gain is time-varying with the amplitude and phase, demonstrating the fading characteristics.



Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System, Fig. 3 Overview of the measurement area at the campus of Shanghai Jiao Tong University, China. A spaced rectangular antenna array with cross polarized patch

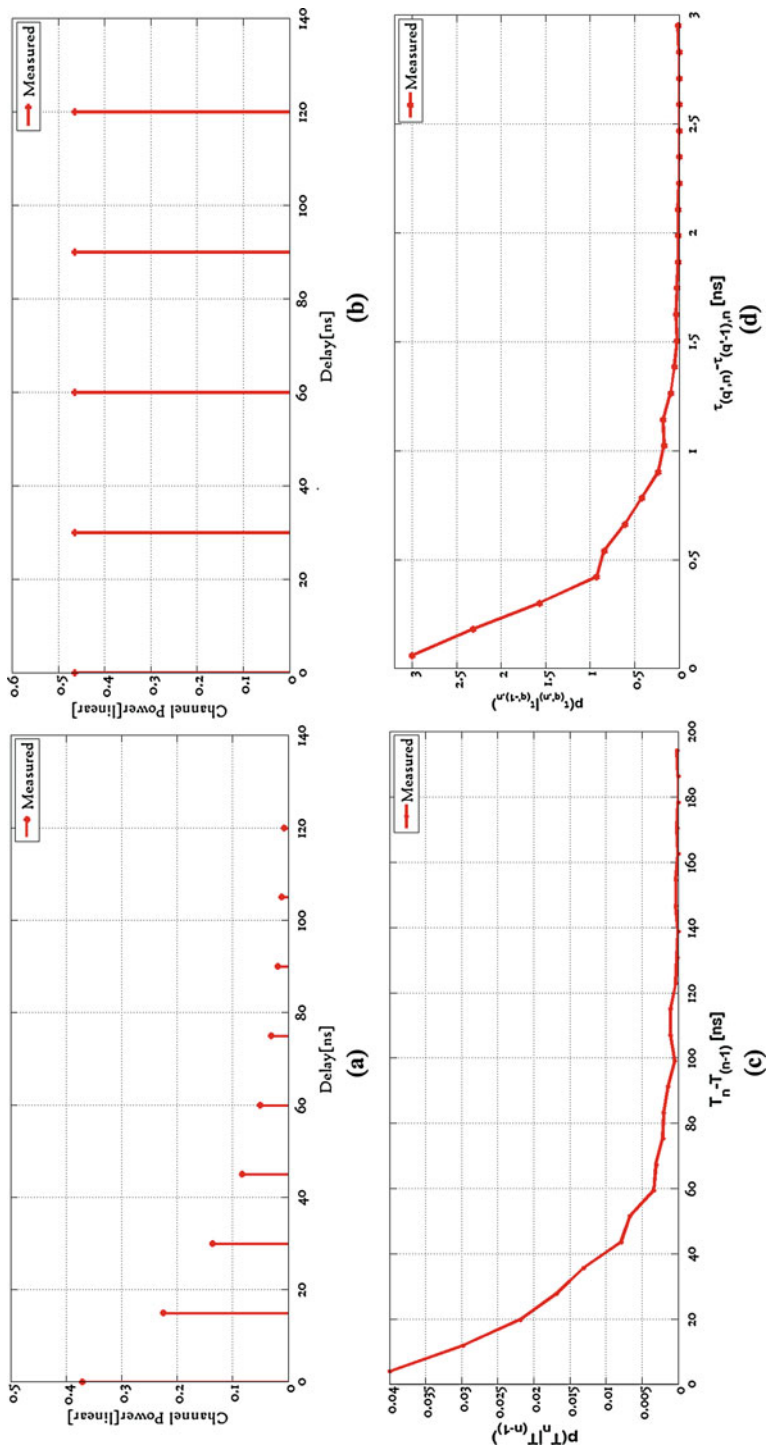
antenna elements at the TX was placed on the rooftop of the biomedical building during their respective measurement campaigns and the spaced uniform linear array as users were located inside the SEIEE building acting as multiple-antenna user



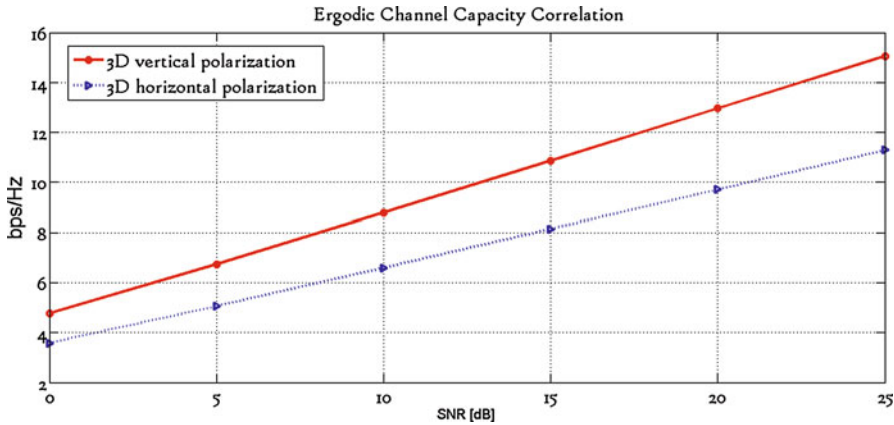
Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System, Fig. 4 Time-domain channel characteristic

In conventional indoor MIMO channel models, the total paths are not sufficient to represent the properties of the reflected cluster, because of the noncooperative situation in elevation angles as well as their properties like delay and capacity correlation which should be considered and developed when calculating channel coefficient as shown in Section “A 3D Massive MIMO for Outdoor-to-Indoor Scenario”

In addition, the PDP, delay properties of arriving cluster and Ray paths of the proposed model over all snapshots, has been investigated. Figure 5a, b generated the corresponding delay on 3D channel model. The results were obtained by different Ray paths, one for a direct path with zero delay $\tau_1 = 0$ and the other paths which are reflection with the delay of $\tau_2 > 0$, each of them has the same power. On the other hand, Figure 5b



Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System, Fig. 5 (a) Ray path delay profile, (b) average channel power, (c) distribution of cluster arrival times, and (d) distribution of Ray arrival times



Three-Dimensional Channel Measurement for the Indoor 5G Multiuser Massive MIMO System, Fig. 6 Ergodic channel capacity correlation

shows an average channel power decreases exponentially with the channel delay. A plot indicates how a transmitted pulse decreases based on the PDP as it will travel through a multipath channel with different propagation delays. In addition, it is illustrated in Figs 5c, d that the delay profile of Ray arrival time with q_{th} Ray path of the $\tau_{q,n}$ in the n_{th} cluster and q_{th} Ray path is generated and that each of them has an exponential distribution of Eqs. (25) and (26). Furthermore, in the plot, the signal power of each multipath is plotted against their respective propagation delays, and the different propagation path lengths cause different propagation time delays. In addition, there might be some signals arriving at a certain delay, but after some time, the power significantly shuts down. If we look at the last arriving paths, they have very low power. They might be insignificant in the instance that the maximum DS is huge. They are very insignificant that it would not be well to consider such paths of low power in the DS.

Finally, we described the ergodic capacity correlation which resulted from the existing 3D channel model that was obtained from Eqs. (32) and (33). We focused and analyzed the ergodic capacity correlation of the RXs including the effect of antenna polarization based on practical multipath wireless communication environment obtained by Sect. IV. The correlation matrix for multiple antenna techniques has been defined

for the evaluation of the performance of the 3D MU-massive MIMO. Figure 6 shows the plot of ergodic correlation capacity between two antenna elements which are vertically polarized and another switching to horizontal polarization because the user movements in different directions have significant different capacity. It is illustrated that the antenna polarization mismatching with the direction of Ray paths is degrading the received channel capacity from time to time. Intuitively, the distance between two inter-elements are at maximum, while the correlation remains minimal. In other words, the channel capacity is varying based on the space between antenna elements increases.

Conclusion

In this entry, we proposed a 3D practical outdoor-to-indoor MU-massive MIMO by investigating the multipath delay including channel correlation and capacity, around Shanghai Jiao Tong University. This consists of defining a composite channel coefficient that is the scaled sums of 3D components were jointly calculated the azimuth and elevation angles for the proposed model. First, we modeled our outdoor channel coefficient and then Ray path model proposed from the reflected clusters which obtained different elevation/azimuth angles and phase excitations.

In addition, their impact on 3D MU-massive MIMO has been studied via statistical properties including received spatial correlation among the massive MIMO antennas. Furthermore, we investigated the essential channel characteristics, such as DS and PDP for different clusters and Ray paths. Finally, ergodic correlation capacity was computed as a function of correlation for multiple RX antennas.

References

- 3GPP (July, 2015) Study on 3d channel model for lte. In: Technical report. 3GPP 36.873 (V12.2.0), 3GPP
- Aghaiezhadfirouzja S, Liu H, Xia B, Tao M (2016) Implementation and measurement of single user MIMO testbed for td-lte-a downlink channels. In: IEEE international conference on communication software and networks, pp 211–215
- Aghaiezhadfirouzja S, Liu H, Xia B, Luo Q, Guo W (2017) Third dimension for measurement of multi user massive MIMO channels based on lte advanced downlink. In: Signal and information processing, pp 758–762
- Cho YS, Kim J, Yang WY, Kang CG (2010) MIMO-OFDM wireless communications with MATLAB. Wiley, Singapore City
- Dao MT, Nguyen VA, Im YT, Park SO, Yoon G (2011) 3d polarized channel modeling and performance comparison of MIMO antenna configurations with different polarizations. IEEE Trans Antennas Propag 59(7):2672–2682
- Gao X, Tufvesson F, Edfors O, Rusek F (2012) Measured propagation characteristics for very-large MIMO at 2.6 GHz. In: Signals, systems and computers, pp 295–299
- Haneda K, Tian L, Asplund H, Li J, Wang Y, Steer D, Li C, Balercia T, Lee S, Kim Y (2016a) Indoor 5g 3gpp-like channel models for office and shopping mall environments
- Haneda K, Zhang J, Tan L, Liu G, Zheng Y, Asplund H, Li J, Wang Y, Steer D, Li C (2016b) 5G 3GPP-like channel models for outdoor urban microcellular and macrocellular environments. In: Vehicular technology conference, pp 1–7
- Kafle PL, Intarapanich A, Sesay AB, McRory J, Davies RJ (2008) Spatial correlation and capacity measurements for wideband MIMO channels in indoor office environment. IEEE Trans Wirel Commun 7(5):1560–1571
- Knudsen MB, Pedersen GF (2002) Spherical outdoor to indoor power spectrum model at the mobile terminal. IEEE J Sel Areas Commun 20(6):1156–1169
- Kyritsi P, Cox DC, Valenzuela RA, Wolniansky PW (2003) Correlation analysis based on MIMO channel measurements in an indoor environment. IEEE J Sel Areas Commun 21(5):713–720
- Payami S, Tufvesson F (2012) Channel measurements and analysis for very large array systems at 2.6 GHz. In: European conference on antennas and propagation, pp 433–437
- Renaudin O, Kolmonen VM, Vainikainen P, Oestges C (2010) Non-stationary narrowband MIMO inter-vehicle channel characterization in the 5-ghz band. IEEE Trans Veh Technol 59(4):2007–2015
- Shafi M, Zhang M, Moustakas AL, Smith PJ, Molisch AF, Tufvesson F, Simon SH (2006) Polarized MIMO channels in 3-d: models, measurements and mutual information. IEEE J Sel Areas Commun 24(3):514–527
- Wang C, Murch RD (2006) Adaptive downlink multi-user MIMO wireless systems for correlated channels with imperfect CSI. IEEE Trans Wirel Commun 5(9):2435–2446
- Weng J, Tu X, Lai Z, Salous S, Zhang J (2014) Indoor massive MIMO channel modelling using ray-launching simulation. Int J Antennas Propag 2014:1
- Wu S, Wang CX, Aggoune EHM, Alwakeel MM, He Y (2014) A non-stationary 3-d wideband twin-cluster model for 5g massive MIMO channels. IEEE J Sel Areas Commun 32(6):1207–1218
- Wu S, Wang CX, Haas H, Aggoune EHM, Alwakeel MM, Ai B (2015) A non-stationary wideband channel model for massive MIMO communication systems. IEEE Trans Wirel Commun 14(3):1434–1446
- Xu H, Chizhik D, Huang H, Valenzuela R (2004) A generalized space-time multiple-input multiple-output (MIMO) channel model. IEEE Trans Wirel Commun 3(3):966–975
- Yuan Y, Wang CX, He Y, Alwakeel MM et al (2015) 3D wideband non-stationary geometry-based stochastic models for non-isotropic MIMO vehicle-to-vehicle channels. IEEE Trans Wirel Commun 14(12):6883–6895

Time and Frequency Synchronization in OFDM Systems

Weile Zhang¹ and Hai Lin²

¹Department of Information and communication Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China

²Department of Electrical and Electronics Systems, Osaka Prefecture University, Sakai, Osaka, Japan

Synonyms

OFDM synchronization

Definitions

Synchronization in OFDM systems means finding the timing and carrier frequency offset of observed OFDM signal.

Historical Background

Orthogonal frequency-division multiplexing (OFDM) has been widely adopted as a mature technique owing to its high data rate transmission capability and robustness to multipath delay spread. Time and frequency synchronization is of greatest importance in OFDM systems. Note that synchronization is an essential task for any digital communication system and required for reliable reception of the transmitted signal.

In general, a wireless receiver does not have prior knowledge of the physical channel or propagation delay associated with the signal. Moreover, to keep the cost low, transceiver devices usually use low cost oscillators which inherently have some frequency offsets. It is well known that, OFDM is extremely sensitive to timing errors and carrier frequency offsets between transceivers. Timing error results in interblock interference (IBI) and must be counteracted to avoid severe performance degradations. On the other hand, frequency offset destroys orthogonality among subcarriers and produces interchannel interference (ICI). Synchronization of an OFDM signal requires finding the block timing and compensating for carrier frequency offset.

The first attempts on OFDM synchronization were reported around 20 years ago. For example, Moose derived the maximum likelihood estimator for carrier frequency offset which is calculated in frequency domain (Moose 1994). Beek et al. described a method using correlation with cyclic prefix (CP) to find the timing point (van de Beek et al. 1995). Schmidl and Cox (1997) presented a classic synchronization algorithm using two repeated training halves, referred to as S&C method. Motivated by these pioneer works, extensive studies have been further reported during the following years:

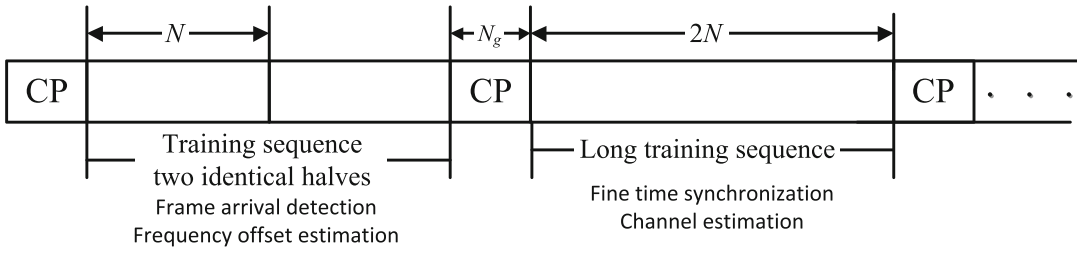
Regarding the time synchronization, the preamble structure of S&C method was modified to obtain a sharper peak in Minn et al. (2003) and Park et al. (2003). Minn et al. proposed an improved algorithm for S&C method to eliminate the influence of CP by introducing difference into the two halves of the training sequence (Minn et al. 2003). Park et al. introduced a new training sequence structure with conjugate symmetry (Park et al. 2003). The start of frame can be blindly estimated by exploiting the characteristics of CP (Krishnamurthy et al. 2005; Ma et al. 2009; Chin 2012). More recently, a class of new methods was presented based on the idea of high-order statistics to improve time synchronization performance (Ziabari and Shayesteh 2013, 2015; Ziabari et al. 2016).

By using time-domain training sequences (Schmidl and Cox 1997; Morelli and Mengali 2000) or frequency-domain embedded pilots (Cui and Tellambura 2007; Gao et al. 2008), carrier frequency offset can be estimated from the data-aided approaches. Meanwhile, different kinds of blind carrier frequency offset estimation methods have also been developed. For example, the frequency offset can be obtained by exploiting the constant modulus constellations (Ghogho and Swami 2002; Roman and Koivunen 2005), the kurtosis-type criterion (Yao and Giannakis 2005), null subcarriers (Liu and Tureli 1998; Chen 2002), or the multi-antenna redundancy (Zhang and Yin 2013).

In addition to the above works, time and frequency synchronization techniques are sometimes considered in conjunction with radio frequency (RF) front-end impairments. For example, frequency synchronization issue has been considered in Lin et al. (2010) under the distortion of I/Q imbalance, i.e., the mismatch between the in-phase and quadrature branches. The effect of oscillator phase noise has been taken into account in Salim et al. (2014).

Foundations

Consider a packet-switched radio communication system with a random access protocol. This



Time and Frequency Synchronization in OFDM Systems, Fig. 1 A typical preamble structure of an OFDM system

essentially means that a receiver has no a priori knowledge about packet arrival time. Synchronization should be completed shortly after the start of the reception of a packet. Hence, the data packet is preceded with a known sequence (the so-called preamble). The preamble should be carefully designed to provide enough information for a good time and frequency synchronization, as well as the subsequent channel estimation. In the following, we adopt the preamble structure as depicted in Fig. 1 and present the typical approaches of time and frequency synchronization for OFDM. Here, we only consider the synchronization in the standard single-user single-antenna OFDM system. For more complicated scenarios such as multi-antenna OFDM, multi-user OFDM, massive multi-input multi-output (MIMO) OFDM, we refer readers to Zhang et al. (2012, 2016, 2018), Zhang and Gao (2016) and the references therein.

Denote Δf as the frequency offset in Hz between the transceivers. Suppose the size of fast Fourier transformation (FFT) window is $2N$. Denote T as the block duration corresponding to the length of FFT window. The sampling interval in system is $T/(2N)$. The baseband discrete-time channel response between transceivers is modeled by a finite impulse filter (FIR), denoted by $h(n)$. Considering the transmitter sends baseband signal $x(n)$, the received baseband samples at the receiver can be expressed as

$$r(n) = (x(n) * h(n)) e^{j\pi n T \Delta f / N} + w(n) \quad (1)$$

where $*$ denotes the linear convolution, and $w(n)$ is the corresponding additive white Gaussian noise (AWGN).

1. Coarse Timing:

Coarse timing synchronization, sometimes referred to as frame arrival detection, means detecting the arrival of burst data frame and determining the approximate starting position of data frame. As illustrated in Fig. 1, the frame can be detected based on a training sequence consisting of two identical halves in time domain (Schmidl and Cox 1997). The training sequence is appended before the data transmission with CP and would remain identical after passing through propagation channel in the absence of noise. The effect of channel distortion can be eliminated via multiplying the conjugate of the sample points of the first half by the corresponding samples of the second half. At the start of frame, the phase of each product pair can be aligned and thus the magnitude of the sum attains its maximum.

Specifically, suppose the training sequence for frame detection is of length $2N$. Given d a starting point in a sliding window of $2N$ samples. Once the sliding window is free of ISI, there holds $r(d+n+N) = r(d+n)e^{j\pi T \Delta f}$, $n = 0, 1, \dots, N-1$, in the absence of noise. This indicates that the first half is identical to the second half (in time order), except for a phase shift caused by the carrier frequency offset. Then, the sum of the product pairs can be expressed as (Schmidl and Cox 1997)

$$P(d) = \sum_{n=0}^{N-1} r^*(d+n) r(d+n+N), \quad (2)$$

The above window slides along in time at the receiver to detect the arrival of frame. The received energy for the first and second halves in

the sliding window can be respectively calculated as

$$R_1(d) = \sum_{n=0}^{N-1} |r(d+n)|^2, \quad (3)$$

$$R_2(d) = \sum_{n=0}^{N-1} |r(d+n+N)|^2. \quad (4)$$

The timing metric function can be defined by

$$Q(d) = \frac{|P(d)|^2}{R_1(d)R_2(d)}, \quad (5)$$

which represents the magnitude of correlation coefficient between the first and second halves in the sliding window. The metric in Eq. (5) can be called as *delayed autocorrelation*. In the presence of frame arrival, the timing metric will reach a plateau when there is no IBI within the sliding window. In frequency selective fading channel, the length of timing metric plateau equals CP length minus the channel length. In practice, since the receiver has no idea whether there is an ongoing transmission, the sliding timing metric can be compared to a predefined threshold. The timing metric larger than the threshold indicates a frame arrival. Otherwise, no frame arrival is detected.

2. Frequency Synchronization:

Once the frame arrival is detected, the optimal point with maximal timing metric

$Q(d)$ can be found by

$$\tau_{FD} = \arg \max_d Q(d), \quad (6)$$

which produces the highest correlation between the first and second halves in the sliding window. Regarding this optimal point, without effect of noise there should hold

$$P(\tau_{FD}) = \left(\sum_{n=0}^{N-1} |r(\tau_{FD} + n)|^2 \right) e^{j\pi T \Delta f}. \quad (7)$$

If $|T \Delta f|$ can be guaranteed to be less than one, the carrier frequency offset estimation can be given by

$$\Delta \hat{f} = \frac{\angle [P(\tau_{FD})]}{\pi T}. \quad (8)$$

The maximum estimated frequency offset is limited, since the angle $\pi T \Delta f$ that can be estimated without phase ambiguity is limited to $\pm\pi$. This relates to a maximum frequency offset of $\Delta f_{\max} = 1/T$, which equals the subcarrier spacing. Nevertheless, a larger range can be achieved by slightly changing the preamble (Zelst and Schenk 2004).

3. Fine Timing:

The fine timing in an OFDM system decides exactly where to place the start of FFT window within the OFDM block. Although OFDM system exhibits a guard interval, making it somewhat robust against timing offsets, a nonoptimal fine timing may cause IBI. Since the delayed autocorrelation in frame detection exhibits a certain plateau with low timing resolution, the optimal point τ_{FD} obtained from Eq. (6) may not provide satisfactory fine timing performance. As depicted in Fig. 1, a more reliable fine timing synchronization can be achieved based on a long training sequence (Zelst and Schenk 2004).

Specifically, denote the time-domain samples in the long training sequence by $c(n)$, $n = 0, 1, \dots, 2N - 1$. The optimal fine timing point can be obtained through correlation of the received signals with $c(n)$. Denote

$$\eta(\tau) = \left\| \sum_{k=0}^{2N-1} r(\tau + k) c^*(k) \right\|^2 \quad (9)$$

which represents the power contribution of the channel impulse response at point τ . The estimated power is window integrated over the CP length N_g . The fine timing is then found by searching the maximum of the window integral (Zelst and Schenk 2004):

$$\tau_{\text{est}} = \arg \max_{\tau} \sum_{l=0}^{N_g-1} \eta(\tau + l). \quad (10)$$

4. Phase Tracking:

The aforementioned processing takes care of the initial synchronization of OFDM receiver. However, any practical frequency offset estimation cannot be perfect. After the initial frequency synchronization, the residual frequency offset would lead to the accumulative phase offset over the subsequent data blocks and result in large detection error rate. Besides, the system may also experience phase noise invoked by the combination of RF oscillator and phase-locked loop (PLL). Both residual frequency offset and PN cause a common phase turn for all carriers, called common phase error (CPE) and a Gaussian-like ICI term (Zelst and Schenk 2004).

It is necessary to estimate and correct the rotation of the received constellation points due to the CPE. A convenient method is to use pilot subcarriers inserted in the data blocks, i.e., subcarriers containing known data, to complete the *phase tracking*. Specifically, the rotation for the i th block can be estimated by

$$\vartheta_i = \angle \left(\sum_{k \in \mathcal{P}} \hat{s}_i(k) s_i^*(k) \right) \quad (11)$$

where $\mathcal{P}$ denotes the set of pilot subcarrier numbers, and $\hat{s}_i(k)$ is the estimate for the pilot symbol $s_i(k)$ after channel equalization. This rotation can be corrected by multiplying the complex symbols on all subcarriers of the i th block with $\exp(-j\vartheta_i)$.

Key Applications

Time and frequency synchronization is an essential task for any digital communication system and required for reliable reception of the transmitted signal.

References

- Chen B (2002) Maximum likelihood estimation of OFDM carrier frequency offset. *IEEE Signal Process Lett* 9(4):123–126
- Chin W-L (2012) Blind symbol synchronization for OFDM systems using cyclic prefix in time-variant and long-echo fading channels. *IEEE Trans Veh Tech* 61(1):185–195
- Cui T, Tellambura C (2007) Joint frequency offset and channel estimation for OFDM systems using pilot symbols and virtual carriers. *IEEE Trans Wirel Commun* 6(4):1193–1202
- Gao F, Cui T, Nallanathan A (2008) Scattered pilots and virtual carriers based frequency offset tracking for OFDM systems: algorithms, identifiability, and performance analysis. *IEEE Trans Commun* 56(4):619–629
- Ghogho M, Swami A (2002) Blind frequency-offset estimator for OFDM systems transmitting constant-modulus symbols. *IEEE Commun Lett* 6(8):343–345
- Krishnamurthy V, Athaudage CRN, Huang D (2005) Adaptive OFDM synchronization algorithms based on discrete stochastic approximation. *IEEE Trans Signal Process* 53(4):1561–1574
- Lin H, Zhu X, Yamashita K (2010) Low-complexity pilot-aided compensation for carrier frequency offset and I/Q imbalance. *IEEE Trans Commun* 58(2):448–452
- Liu H, Tureli U (1998) A high-efficiency carrier estimator for OFDM communications. *IEEE Commun Lett* 2(4):104–106
- Ma S, Pan X, Yang G-H, Ng T-S (2009) Blind symbol synchronization based on cyclic prefix for OFDM systems. *IEEE Trans Veh Tech* 58(4):1746–1751
- Minn H, Bhargava VK, Letaief KB (2003) A robust timing and frequency synchronization for OFDM systems. *IEEE Trans Wirel Commun* 2(4):822–839
- Moose P (1994) A technique for orthogonal frequency division multiplexing frequency offset correction. *IEEE Trans Commun* 42(10):2908–2914
- Morelli M, Mengali U (2000) Carrier-frequency estimation for transmissions over selective channels. *IEEE Trans Commun* 48(9):1580–1589
- Park B, Cheon H, Kang C, Hong D (2003) A novel timing estimation method for OFDM systems. *IEEE Commun Lett* 7(5):239–241
- Roman T, Koivunen V (2005) Subspace method for blind CFO estimation for OFDM systems with constant modulus constellations. In: *Proceeding of the IEEE Vehicular Tech. Conf.*, pp 1253–1257
- Salim OH, Nasir AA, Mehrpouyan H et al (2014) Channel, phase noise, and frequency offset in OFDM systems: joint estimation, data detection, and hybrid cramer-rao lower bound. *IEEE Trans Commun* 62(9):3311–3325
- Schmidl TM, Cox DC (1997) Robust frequency and timing synchronization for OFDM. *IEEE Trans Commun* 45(12):1613–1621
- Beek J-Jvan de, Sandell M, Isaksson M, Borjesson P (1995) Low complex frame synchronization in OFDM systems. In: *Proceedings of ICUPC*, pp 982–986

- Yao Y, Giannakis GB (2005) Blind carrier frequency offset estimation in SISO, MIMO, and multiuser OFDM systems. *IEEE Trans Commun* 53(1):173–183
- Zelst AV, Schenk TCW (2004) Implementation of a MIMO OFDM-based wireless LAN system. *IEEE Trans Signal Process* 52(2):483–494
- Zhang W, Gao F (2016) Blind frequency synchronization for multiuser OFDM uplink with large number of receive antennas. *IEEE Trans Signal Process* 64(9):2255–2268
- Zhang W, Yin Q (2013) Blind maximum likelihood carrier frequency offset estimation for OFDM with multi-antenna receiver. *IEEE Trans Signal Process* 61(9):2295–2307
- Zhang W, Gao F, Yin Q, Nallanathan A (2012) Blind carrier frequency offset estimation for interleaved OFDMA uplink. *IEEE Trans Signal Process* 60(7):3616–3627
- Zhang W, Yin Q, Gao F (2016) Computationally efficient blind estimation of carrier frequency offset for MIMO-OFDM systems. *IEEE Trans Wirel Commun* 15(11):7644–7656
- Zhang W, Gao F, Jin S, Lin H (2018) Frequency synchronization for uplink massive MIMO systems. *IEEE Trans Wireless Commun* 17(1):235–249
- Ziabari HA, Shayesteh MG (2013) Sufficient statistics, classification, and a novel approach for frame detection in OFDM systems. *IEEE Trans Veh Tech* 62(6):2481–2495
- Ziabari HA, Shayesteh MG (2015) Novel coarse timing synchronization methods in OFDM systems using fourth-order statistics. *IEEE Trans. Veh. Tech* 64(5):1904–1917
- Ziabari HA, Zhu W-P, Swamy MNS (2016) Improved coarse timing estimation in OFDM systems using high-order statistics. *IEEE Trans Commun* 64(12):5239–5253

Time-Domain Synchronous Orthogonal Frequency Division Multiplexing

Jian Song

Tsinghua University, Beijing, China

Definition

One of the padding schemes for orthogonal frequency division multiplexing (OFDM) system uses the training sequence instead of conventional cyclic-prefix (CP) or all-zero padding (ZP) as the guard interval between two adjacent OFDM symbols in time domain. The advantage of using the known sequence as the guard interval (i.e., padding) is that the system synchronization can be done in the time domain instead of in the frequency domain like C-OFDM, and that is why it is called Time-Domain Synchronous Orthogonal Frequency Division Multiplexing (TDS-OFDM).

Historical Background

The TDS-OFDM scheme was first proposed to be utilized in the Chinese Digital Television Terrestrial Broadcasting (DTTB) system with the full name of Digital Television Terrestrial Multimedia Broadcasting (DTMB) (Song et al. 2007).

Time Series

- [Data Analytics for Longitudinal Biomedical Data](#)

Time Synchronization of Nanomachines

- [Synchronization of Nanomachines](#)

Foundations

Orthogonal Frequency Division Multiplexing (OFDM) is a kind of multicarrier modulation scheme which is developed from the traditional frequency division multiplexing (FDM) scheme. The major difference between OFDM and FDM systems is that all subcarrier signals within OFDM system are orthogonal to each other, and there is no guard band needed which provides better spectrum utilization. OFDM is widely utilized for both wireless communication systems and digital broadcasting systems to effectively

and efficiently cope with the very harsh wireless transmission environment with severe multipath fading.

When OFDM scheme is utilized, the entire wideband signal can be considered as the superposition of many synchronized subcarrier signals. If the frequency spacing between two adjacent subcarriers is narrow enough so that the subchannel frequency response can be taken as a constant, the channel equalization could be quite simple and easy to be implemented.

To effectively deal with the inter (OFDM) symbol interference (ISI) caused by the multipath effect of the channel in the time domain, guard interval (i.e., padding in this paper) is needed between two adjacent OFDM symbols, so that two adjacent OFDM symbols would not overlap in the time domain. Of course, the guard interval must have its length at least equal to the maximum delay spread of the multipath channel. There are three typical padding schemes and they are:

1. Cyclic-prefix (CP)-OFDM system: It prefixes each OFDM symbol with the repetition of its end of certain length. Cyclic prefix not only serves as the guard interval to eliminate ISI effect but also converts the linear convolution of the multipath channel with the input signal into circular convolution. This greatly facilitates the frequency domain processing for the synchronization and channel estimation. In order to estimate the channel frequency response and support the system synchronization, pilots (known symbol sequence carried by subcarriers) are indispensable such as continuous and scattered pilots in the Digital Video Broadcasting specification (ETSI Standard EN 300 744 v1.6.2 2015).
2. Zero padding (ZP)-OFDM system: Zero symbols are appended to each OFDM symbol in the time domain. At the cost of the increased receiver complexity (the single FFT required by CP-OFDM will be replaced by FIR filtering), ZP-OFDM guarantees symbol recovery and assures FIR (even zero-forcing) equalization of FIR channels regardless of the channel zero locations (Muquet et al. 2002). Pilots are

also needed for the system synchronization and channel estimation.

3. Time-Domain Synchronous (TDS)-OFDM system: Similar to ZP-OFDM system, certain amount of time-domain symbols carrying the known sequence is prefixed to each OFDM symbol. The system synchronization can be easily and quickly achieved by the correlation operation in time domain (Song et al. 2007). The sequence prefixed to each OFDM symbol is not necessarily identical and different sequence can be taken as the unique identification of OFDM symbol which facilitates the multiple services support (i.e., all the OFDM symbols belonging to one service can be assigned the same sequence while different services will have different known sequences). ZP-OFDM may be considered a special case of TDS-OFDM with the known sequence of zeros.

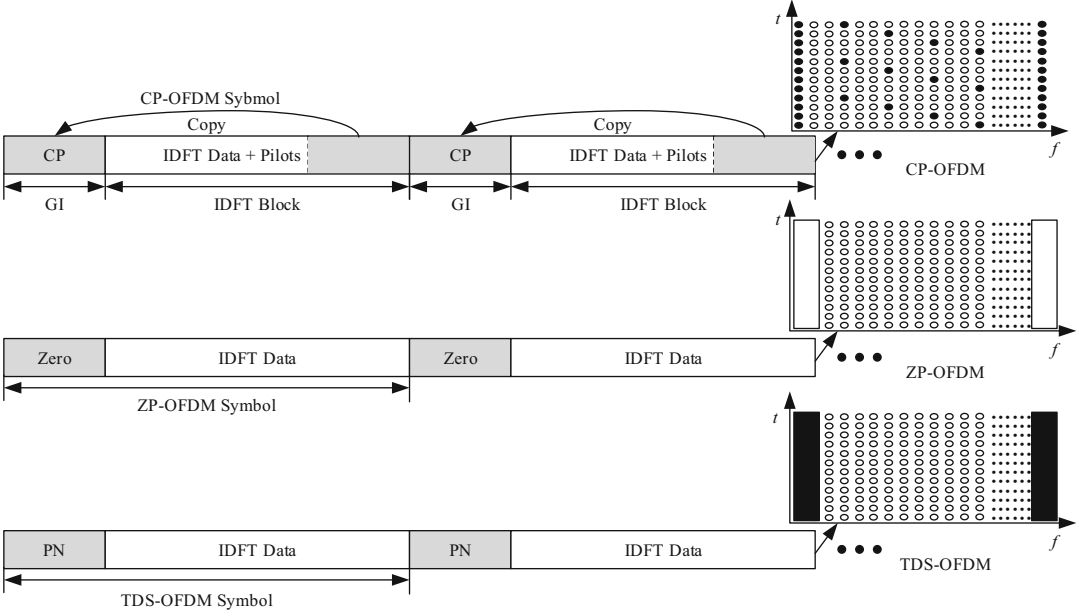
The constitution of three OFDM padding schemes is illustrated in Fig. 1.

The spectrum efficiencies of the three OFDM padding schemes without considering the overhead for channel estimation are identical if the length of the guard interval is the same. However, as mentioned previously, the spectrum efficiency of CP- and ZP-OFDM will be further reduced due to the utilization of pilots.

For TDS-OFDM scheme, the sequence could be of any type depending on the design requirement.

Using the frame structure of DTMB as one example, the basic building block is the signal frame which consists of known sequence as the frame header and OFDM symbol as frame body, as shown in Fig. 2.

The pseudonoise (PN) sequences used in frame header mode PN420 is defined as the cyclically extended 8th-order M-sequence and implemented by Fibonacci Type linear feedback shift register (LFSR). Each PN sequence bit will be mapped to a nonreturn to zero binary signal by mapping zero to +1 and one to -1. To achieve better synchronization and channel estimation, the power of frame header is twice as much as that of frame body.



Time-Domain Synchronous Orthogonal Frequency Division Multiplexing, Fig. 1 The illustration for three different OFDM padding schemes

Frame header mode PN420 (420 symbols)(55.6μs)	Frame body(System information and data) (3780 symbols)(500μs)
--	--

Time-Domain Synchronous Orthogonal Frequency Division Multiplexing, Fig. 2 Structure of the signal frame in DTMB with frame header mode PN420

Pre-amble: 82 symbols	PN255 Symbols	Post-amble: 83 symbols
--------------------------	---------------	---------------------------

Time-Domain Synchronous Orthogonal Frequency Division Multiplexing, Fig. 3 The structure of PN420

Frame header (dual K-length PN-MC sequences)	Frame body (N symbols)
---	------------------------

Time-Domain Synchronous Orthogonal Frequency Division Multiplexing, Fig. 4 The generic structure of the signal frame in DTMB-A

The frame header mode PN420 consists of a preamble, a 255-bit-long PN sequence (PN255) and a post-amble, as shown in Fig. 3. The pre-ample and post-amble are defined as the cyclical extension of PN255 sequence for a better protection on PN255 which will be used for channel estimation. The frame headers may be different for the different signal frames within the same

super frame (the next level of DTMB frame structure) for identification purpose. For another example, the frame header of signal frame in the advanced version of DTMB called DTMB-A (Song et al. 2015) is slightly different and is shown in Fig. 4. In dual multicarrier PN (PN-MC) structure, the first PN-MC sequence can be regarded as

the cyclic prefix of the second so that better channel estimation for higher constellation mapping can be achieved by eliminating the mutual interference between PN and IDFT data. PN-MC sequence is the IFFT of the frequency-domain binary sequence with constant amplitude in frequency domain to support more accurate channel estimation. The PN-MC sequences in DTMB-A system are carefully selected with lower time-domain peak-to-average power ratio to provide better system performance.

Key Applications

TDS-OFDM has been successfully adopted into DTMB system and also the advanced version of DTMB system called DTMB Advance (DTMB-A). Both systems are now adopted as the International Telecommunication Union standards (Recommendation ITU-R BT. 1306-7 [2015](#)).

References

- ETSI Standard EN 300 744 v1.6.2, Oct. 2015 Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television
- Muquet B, Wang Z, Giannakis GB, de Courville M, Duhamel P (2002) Cyclic-prefixing or zero-padding for wireless multicarrier transmissions? *IEEE Trans Commun* 50(12):2136–2148
- Recommendation ITU-R BT. 1306-7, Jun. 2015 Error-correction, data framing, modulation and emission methods for digital terrestrial television broadcasting
- Song J, Yang Z, Yang L, Gong K, Pan C, Wang J, Wu Y (2007) Technical review on Chinese digital terrestrial television broadcasting standard and measurements on some working modes. *IEEE Trans Broadcast* 53(1):1–7
- Song J, Yang Z, Wang J (2015) Digital television broadcasting technology and system. Wiley-IEEE Press, Piscataway

Tiny Volumes of Fluids

- [Communications and Networking in Droplet-Based Microfluidic Systems](#)

Topology Control

- [Contact Plan Design for Space Disruption/Delay-Tolerant Networks](#)
- [Topology Control in Mobile Ad Hoc Networks with Cooperative Communications](#)

Topology Control in Mobile Ad Hoc Networks with Cooperative Communications

Quansheng Guan

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Synonyms

[Topology control](#); [Topology management](#)

Definitions

Network topology depicts the interconnections of nodes and the connectivity of communication networks. Topology control relates to where to deploy the links and how the link works to form a good network topology. Mobile ad hoc network (MANET) is a dynamic network where nodes establish connections with each other without an infrastructure. A node in MANETs is not only a terminal host that is the source and the destination for its own data flows but also a router that forwards data packets for other flows. Neighbor discovery and topology organization are the important self-organization activities in MANETs. Due to the lack of centralized control, MANETs rely on distributed topology control by nodes, rather than a centralized network controller. Topology control in MANETs refers to as neighboring management that selects a set of its neighbors to dynamically establish data links and adjust transmit power accordingly while preserving network

connectivity. In this sense, topology control has extended the link-wide neighbor management to the network-wide topology management.

Historical Background and Key Applications

Network topology has significant impacts on network performance, such as energy consumption, reliability, and capacity.

Topology control is first adopted in wireless sensor networks (WSNs) to save energy in nodes and prolong the lifetime of the network, since energy consumption is the critical issue in WSNs (Santi 2005). A long transmission range corresponds to a high transmission power. Topology control is to minimize the critical transmission range of each node (Santi and Blough 2003) while preserving network connectivity. In this sense, the energy consumption is reduced, and the network lifetime is prolonged. To achieve distributed topology control, some works try to study the minimize node degree (Moraes et al. 2009; Guan et al. 2009), i.e., the number of neighboring nodes, that can preserve network connectivity. The work in Guan et al. (2010) preserves the long-duration links in network topology to eliminate the impact of intermittent connections in MANETs. From the reliability perspective, nodes prefer transmitting at their maximum radio power to have more neighbors. However, interference may be introduced due to the long link distance, probably leading to less throughput per node. The works in Guan et al. (2008, 2011) take into consideration the interference to achieve capacity-optimized network topology. Other works study the security issues in topology control in Galitos (2008) and Guan et al. (2012b).

Topology control in traditional MANETs is carried out in complex networks of simple links, which attempts to create, adapt, and manage a network on a maze of point-point noncooperative wireless links. Cooperative communication exploits user diversity to increase spectral and power efficiency and reduce outage probability in Woradit et al. (2009) and Scaglione and Hong

(2003). The emerging cooperative communications, which provide more flexibilities in communication manners, have enabled simple networks with complex links in Guan et al. (2012a).

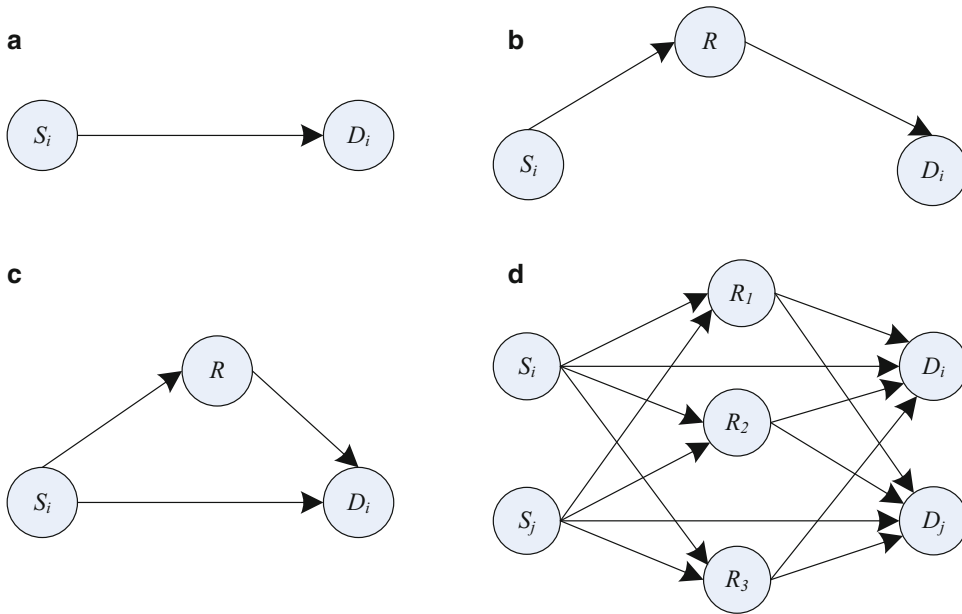
Topology Control in Mobile Ad Hoc Networks with Cooperative Communications

Cooperative communications often involve the participation of multiple nodes, which complicates the links and neighborhood relationship. Cooperative communication, as a physical-layer technique, has a significant impact on upper-layer issues like topology control and thus gives rise to another dimension of challenges to topology control in MANETs.

Cooperative Communications

With the introduction of cooperative communications to MANETs, each link may work in one of the following transmission manners as shown in Fig. 1:

1. *Direct transmission*: The upstream source transmits its data packets directly to its next-hop destination. No relay node is involved in direct transmissions.
2. *Two-hop transmission*: The upstream source's data packets are forwarded to the next-hop destination via two short links rather than the direct long link, with the assistance of a nearby neighbor.
3. *Relay transmission*: The source first broadcasts its data packets to a relay and the destination in the first slot, while the relay forwards the received data packets to the destination in the second slot. The signals from the source and the relay are combined to decode the source's information. The relay can decode and forward the source's signal, or just amplify and forward the source's signal, to the destination.
4. *Interference cancellation-based transmission*: The source (e.g., S1 in Fig. 1) first selects another cooperative source (e.g., S2 in Fig. 1) and three assisting relays (e.g., R1, R2, and



Topology Control in Mobile Ad Hoc Networks with Cooperative Communications, Fig. 1 Transmission patterns. (a) DT transmission. (b) TT transmission. (c) DF transmission. (d) IC transmission

R3 in Fig. 1). S1 and S2 broadcast their packets to the three relays concurrently in the first slot. Each relay then scales the received signals and forwards them to the destinations concurrently. In contrast to the traditional wisdom of separating concurrent transmissions far away to avoid mutual interference, interference cancellation-based transmissions allows simultaneous multiple source-destination transmissions by canceling the mutual interference with the help of multiple cooperative relays.

Topology Control with Cooperative Communications

Link-level physical-layer issues are the main focus of most existing works on cooperative communications, such as outage probability and outage capacity in Woradit et al. (2009), Seddik et al. (2006), and Ding et al. (2011). However, the introduction of cooperative communications has complicated the link management in MANETs.

Nodes in MANETs face different space distributions and interference environment. Different transmission manners may achieve different

interference levels, energy efficiencies, as well as cooperation diversities (Li et al. 2011; Ao et al. 2014; Guo et al. 2014). The nodes in MANETs will select a suitable transmission manner based on their observed conditions. The selection of transmission manners thus leads to *selective diversity* for each transmission. Hence, the link management in MANETs with cooperative communications simplifies to a diversity selection issue. Topology control with complex links of selective diversity then forms a new efficient topology. The nodes who form a cooperative coalition establish a supper link.

An Energy-Efficient Topology Control

This section will illustrate an example of energy-efficient topology control.

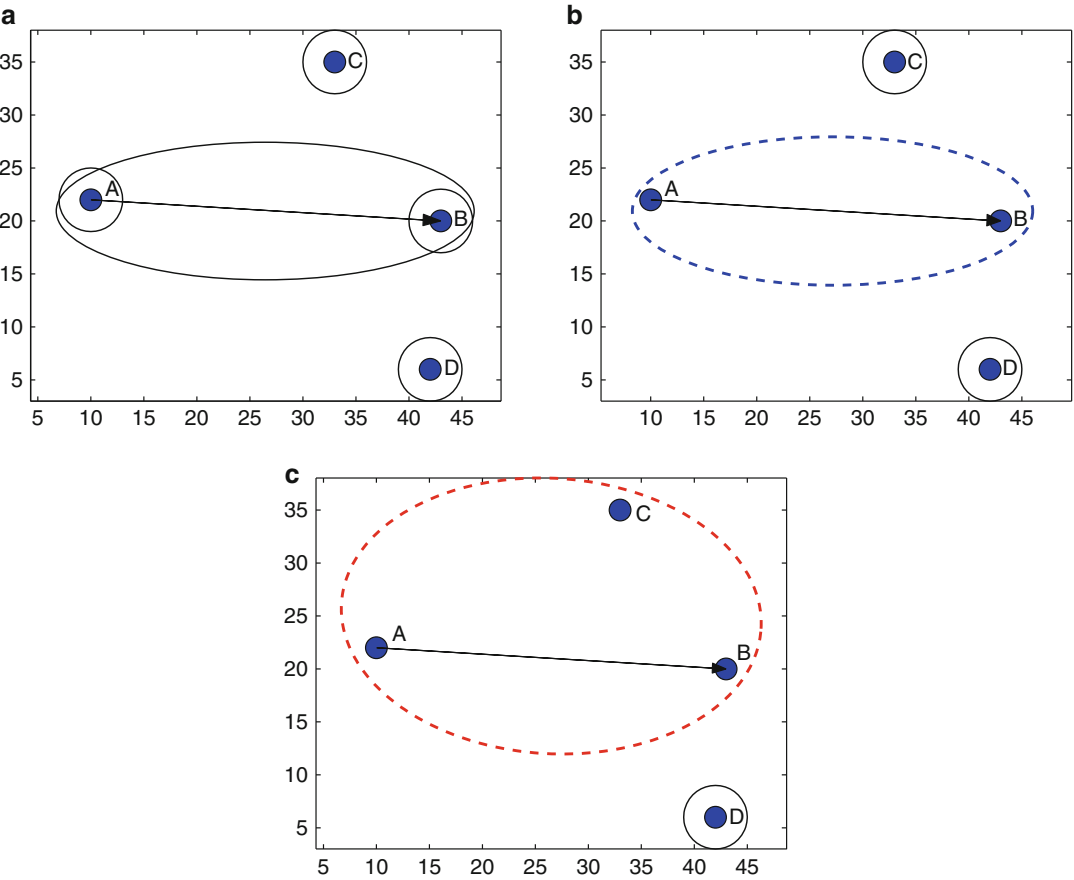
Energy efficiency refers to as the achievable information transmission per Joule energy consumption with bits per Joule as the unit. Obviously, interference cancellation-based transmission has advantages on information transmission since it allows simultaneous transmissions. However, it also involves more relays in the transmissions, which may reduce the energy efficiency. In

this sense, links need to be established carefully to form an energy efficient topology.

Nodes in MANETs select the cooperative patterns for transmissions. A node may choose to work in relay transmissions or choose to collaborate with other users to work in interference cancellation-based transmissions. Each node can work in only one transmission pattern. They compete to establish their energy-efficient links selfishly and distributedly, which deserves an analysis from a game-theoretic perspective. Using the energy efficiency as the coalition value, nodes will reach a consensus in transmission pattern selection by an adaptive coalition formation approach which is based on merge and split rules. The coalitions are split or

merged according the resulting coalition value, i.e., the energy efficiency of the formed coalition. The split and merge rules will converge and result in an energy-efficient topology.

The example in Fig. 2 is used to illustrate how the split and merge rules converge to energy-efficient coalitions. In this example, there are four nodes and a transmission pair from node A to node B. In the initialization phase as shown in Fig. 2a, five coalitions are formed, four of which are formed by the four nodes and the last one is formed by the transmission nodes A and B. In the adaptive coalition formation as shown in Fig. 2b, all the coalitions conduct the merge-and-split operations. The coalition formed by node A and the coalition formed by node B



Topology Control in Mobile Ad Hoc Networks with Cooperative Communications, Fig. 2 An example to illustrate coalition formation. (a) Partition initialization.

(b) Adaptive coalition formation. (c) The final network partition

merge into one coalition. After the iteration of adaptive formation based on merge-split rules, with the relay of node C, the transmission can achieve better energy efficiency, and therefore these three nodes form a large coalition as shown in Fig. 2c.

Future Directions

Links in MANETs become more complex when taking into consideration the dynamic spectrum availability and mobility. Topology control in this case goes beyond determining the critical transmission range and selection of transmission patterns.

In practice, topology control often resides between routing and link layers in the protocol stack. Topology control exploits the information from the lower layers, i.e., the link and physical layers, to establish a desirable network topology. The impact of topology control on the upper-layer issues needs further study. For example, routing is carried out over a give topology. The energy-efficient topology may be achieve a desirable end-to-end performance for routing.

Cross-References

► Ad Hoc Networks

References

- Ao X, Yu FR, Jiang S, Guan Q, Leung VC (2014) Distributed cooperative topology control for WANETs with opportunistic interference cancellation. *IEEE Trans Veh Technol* 63(2):789–801
- Ding H, Ge J, Benevides da Costa D, Guo Y (2011) Outage analysis for multiuser two-way relaying in mixed Rayleigh and Rician fading. *IEEE Commun Lett* 15(4):410–412
- Galiotos P (2008) Security-aware topology control for wireless ad-hoc networks. In: *Proceedings of IEEE GLOBECOM'08*, New Orleans, pp 1–6
- Guan Q, Jiang S, Ding QL, Wei G (2008) Impact of topology control on capacity of wireless ad hoc networks. In: *Proceedings of IEEE ICCS*, Guangzhou, pp 588–592
- Guan Q, Ding Q, Jiang S (2009) A minimum energy path topology control algorithm for wireless multihop networks. In: *Proceedings of IWCMC*, Leipzig, pp 557–561
- Guan Q, Yu F, Jiang S, Wei G (2010) Prediction-based topology control and routing in cognitive radio mobile ad hoc networks. *IEEE Trans Veh Technol* 59(9):4443–4452
- Guan Q, Yu FR, Jiang S, Leung VC (2011) Capacity-optimized topology control for MANETs with cooperative communications. *IEEE Trans Wirel Commun* 10(7):2162–2170
- Guan Q, Yu FR, Jiang S, Leung VC (2012a) Topology control in mobile ad hoc networks with cooperative communications. *IEEE Wirel Commun Mag* 19(2):74–79
- Guan Q, Yu FR, Jiang S, Leung VC (2012b) Joint topology and security design in mobile ad hoc networks with cooperative communications. *IEEE Trans Veh Technol* 61(6):2674–2685
- Guo B, Guan Q, Yu FR, Jiang S, Leung VC (2014) Energy-efficient topology control with selective diversity in cooperative wireless ad hoc networks: a game-theoretic approach. *IEEE Trans Wirel Commun* 13(11):6484–6495
- Li H, Jaggi N, Sikdar B (2011) Relay scheduling for cooperative communications in sensor networks with energy harvesting. *IEEE Trans Wirel Commun* 10(9):2918–2928
- Moraes R, Ribeiro C, Duhamel C (2009) Optimal solutions for fault-tolerant topology control in wireless ad hoc networks. *IEEE Trans Wirel Commun* 8(12):5970–5981
- Santi P (2005) Topology control in wireless ad hoc and sensor networks. *ACM Comput Surv (CSUR)* 37(2):164–194
- Santi P, Blough D (2003) The critical transmitting range for connectivity in sparse wireless ad hoc networks. *IEEE Trans Mobile Comput* 2:25–39
- Scaglione A, Hong Y (2003) Opportunistic large arrays: cooperative transmission in wireless multihop ad hoc networks to reach far distances. *IEEE Trans Signal Process* 51(8):2082–2092
- Seddik K, Sadek A, Su W, Liu K (2006) Outage analysis of multi-node amplify-and-forward relay networks. In: *Proceedings of IEEE WCNC'06*, vol 2, Las Vegas, NV, USA
- Woradit K, Quek T, Suwansantisuk W, Win M, Wuttisitkulkij L, Wymeersch H (2009) Outage behavior of selective relaying schemes. *IEEE Trans Wirel Commun* 8(8):3890–3895

Topology Design

► Contact Plan Design for Space Disruption/Delay-Tolerant Networks

Topology Management

- [Topology Control in Mobile Ad Hoc Networks with Cooperative Communications](#)

Traffic Engineering

- [Enhancing QoE-Aware Wireless Edge Caching with Software-Defined Wireless Networks](#)

Traffic Management

- [Wireless Sensor Networks Enabled Urban Traffic Management](#)

Transaction

- [Auction](#)

Transmission Capacity

Steven Weber
Electrical and Computer Engineering
Department, Drexel University, Philadelphia,
PA, USA

Synonyms

[Outage probability](#); [Spectral efficiency](#)

Definition

This entry covers the definition and application of *transmission capacity*, a performance measure for random access wireless networks that, given its tractability when combined with the *stochastic geometric* framework and results in widespread

use in wireless network research, has proven to be fruitful in illuminating the relationships and trade-offs among the various parameters of the network.

Ad hoc and cellular networks. The transmission capacity metric may be profitably applied in the context of both ad hoc (i.e., decentralized, operating without the benefit of fixed infrastructure) and *cellular* (i.e., *with* infrastructure, such as base stations or access points) networks. The key assumption is that at any point in time the non-idle radios in the network have formed into transmitters and receiver pairs, meaning each transmitter has a unique intended receiver, and that receiver only hopes to receive from that transmitter (the so-called *dumbbell model*).

Aloha random access. The transmission capacity metric is best understood as capturing performance in the *medium access* layer of the network protocol stack, i.e., above the physical layer and below the network (routing) layer. In particular, the transmission capacity framework is most often applied in networks employing a random access protocol, specifically, the classical Aloha slotted random access protocol. Under this protocol, each radio having a packet for transmission makes, in each time slot, an independent random decision whether or not to attempt transmission with some fixed probability common to all the radios in the network. This probability is called the random access probability of the network and is viewed as a tunable parameter.

Stochastic geometry. The mathematical performance analysis of wireless networks has been revolutionized by the tractable application of stochastic geometry. We restrict our discussion to the most prevalent (and most tractable) example, namely, the Poisson point process (PPP). A PPP is a countably infinite collection of random points in space obeying the two properties that (i) the random number of points in any finite area set is a Poisson random variable and (ii) the random number of points in any two disjoint finite areas is independent random variables. In the case of transmission capacity, the two key assumptions are that (i) the radios employ the slotted Aloha random access protocol

and (ii) the locations of *all* of the radios in the network at any point in time constitute a realization of a PPP, of a given spatial intensity, i.e., the average number of radios per unit area. Together, these assumptions imply that the locations of the *transmitting* radios form a Poisson process with a reduced intensity (this follows from the “thinning property” of the Poisson process).

SINR. The typical model for the success or failure of a transmission attempt in a wireless model is the so-called *signal to interference plus noise ratio (SINR)* model (also known as the *physical* model). Consider a radio (called the *reference receiver*) seeking to receive a message (called the signal) from a transmitting radio (called the reference transmitter) in the presence of other transmitting radios, each seeking to transmit messages to other radios, but employing the same channel (frequency), so that their messages constitute *interference* from the perspective of the reference receiver. Both the signal message and the interference messages are subject to path loss attenuation, fading, and shadowing as they propagate through the wireless medium. The *channel model* describes the attenuation of the power of these signals as a function of the distance from the origin to the destination. In the typical case, the path loss attenuation over distance d is taken to be $d^{-\alpha}$, where α is the *path loss attenuation constant* and the fading is assumed to be Rayleigh distributed. The SINR model asserts that the attempted transmission between the reference radio pair is successful if the SINR at the reference receiver exceeds a specified threshold.

Outage probability. If the SINR should be lower than this threshold, then the attempted transmission fails, and we say the receiver is in *outage*. The *outage probability* is therefore the probability that a *typical receiver* (located, without loss of generality, at the origin), subject to interference generated by radios whose positions are the realization of a (thinned) Poisson point process, is in outage. There are three separate sources of randomness that determine the outage probability: (i) the random access process that determines which random subset of the network

radios elects to transmit in the current time slot, (ii) the random Poisson point process that determines the locations of those radios, and (iii) the random (Rayleigh) fading that determines the channels connecting the reference transmitter and receiver and the channels connecting each interfering radio with the reference receiver. By the thinning process mentioned above, it is conventional to combine the first two processes.

Definition of transmission capacity. The transmission capacity (TC) is a measure of the maximum spatial intensity of successful concurrent wireless transmissions supportable under a constraint on the outage probability. We now introduce mathematical notation in order to make the definition precise. Let the outage probability constraint parameter be denoted $\varepsilon \in (0, 1)$, and let $\lambda(\varepsilon)$ denote the TC as a function of ε . Let $q(\lambda)$ denote the outage probability as a function of the spatial intensity of the PPP representing the locations of the active transmitters in a time slot, where λ is the spatial intensity parameter for that process. We write $q(\lambda) = \mathbb{P}(\text{SINR}_o(\lambda) < \theta)$ to denote the outage probability, i.e., the probability that the SINR measured at the reference receiver, located at the origin o , is below the specified SINR threshold, θ .

Assuming all radios employ a common transmission power, we may write the SINR as $\text{SINR}_o(\lambda) = u^{-\alpha} h_o / (I_o(\lambda) + \eta)$, where u is the distance separating the reference pair of radios, h_o is the (unit rate) exponential random variable representing the Rayleigh fade on the reference pair's channel, I_o is the random interference seen at the reference receiver (independent of h_o), and η is the inverse signal-to-noise ratio (SNR), i.e., ambient noise normalized by the transmission power.

The (random) interference seen at the reference receiver, $I_o(\lambda)$, may be expressed as the sum of the interferences received from each interfering transmitter, $I_o(\lambda) = \sum_{\mathbf{x}_i \in \Pi(\lambda)} |\mathbf{x}_i|^{-\alpha} h_i$. Here, $\Pi(\lambda) = \{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ is the random PPP, holding the random locations of the interfering radios, and $|\mathbf{x}_i|^{-\alpha}$, h_i represents, respectfully, the path loss attenuation and Rayleigh fading on the channel from transmitter i to the reference receiver at the origin o , of distance $|\mathbf{x}_i|$.

It should be evident that $q(\lambda)$, the outage probability viewed as a function of the intensity of interference, is an increasing function of λ . As an increasing function onto $(0, 1)$, it has an inverse function $q^{-1}(\varepsilon)$, with the defining property that $q(q^{-1}(\varepsilon)) = \varepsilon$, for any $\varepsilon \in (0, 1)$. Note that $q^{-1}(\varepsilon)$ returns the intensity of *attempted* transmissions corresponding to an outage probability of ε . As is intuitive, $q^{-1}(\varepsilon)$ is also increasing: relaxation of the outage constraint results in a higher supportable spatial intensity of attempted transmissions. The transmission capacity, finally, is defined as $\lambda(\varepsilon) \equiv q^{-1}(\varepsilon)(1 - \varepsilon)$, i.e., the intensity of attempted transmissions corresponding to an outage probability of ε , thinned by the success probability $1 - \varepsilon$.

Analysis

Let $d \in \{1, 2, 3\}$ denote network dimension, e.g., $d = 2$ for planar networks, and define the *characteristic exponent* $\delta \equiv d/\alpha$. Leveraging results first established in the seminal paper Baccelli et al. (2006), the outage probability $q(\lambda)$ and transmission capacity $\lambda(\varepsilon)$ under the above model may be computed explicitly, for a known constant c_d :

$$q(\lambda) = 1 - \exp \left\{ -\lambda \frac{\pi \delta c_d}{\sin(\pi \delta)} \theta^\delta u^d - \frac{\theta}{\eta} \right\}$$

$$\lambda(\varepsilon) = \frac{\sin(\pi \delta) \left(-\log(1 - \varepsilon) - \frac{\theta}{\eta} \right) (1 - \varepsilon)}{\pi \delta c_d \theta^\delta u^d} \quad (1)$$

The outage probability and transmission capacity demonstrate the dependence of these performance indicators upon the natural parameters of the network, i.e., the network dimension d , the characteristic exponent $\delta = d/\alpha$, the SINR threshold θ , the reference transmitter receiver separation distance u , and the SNR η , as well as the core relationship between the spatial intensity of attempted transmissions λ and the corresponding outage probability at a typical receiver q .

Historical Background

The transmission capacity metric and its analysis via stochastic geometry were first introduced in Weber et al. (2005). The application of the framework to various settings and design scenarios followed over the subsequent 10 years, including successive interference cancellation (Weber et al. 2007b), fading-based scheduling (Weber et al. 2007a), bandwidth partitioning (Jindal et al. 2008a), power control (Jindal et al. 2008b), multi-hop random access (Andrews et al. 2010), multi-antenna communication (Jindal et al. 2011), two-way communication (Vaze et al. 2011), directional antennas (Wildman et al. 2014), spectrum games (Malanchini et al. 2015), and network model comparison (Wildman and Weber 2018). Both an overview article (Weber et al. 2010) and a monograph (Weber and Andrews 2012) on transmission capacity are available. The recent paper (Kalamkar and Haenggi 2018) argues that a related (but distinct) metric, termed the *spatial outage capacity*, is a more accurate measure of capacity than the transmission capacity.

Cross-References

- [Capacity of Wireless Ad Hoc Networks](#)
- [Network Information Theory](#)

References

- Andrews JG, Weber S, Kountouris M, Haenggi M (2010) Random access transport capacity. *IEEE Trans Wirel Commun* 9(6):2101–2111
- Baccelli F, Blaszcyszyn B, Muhlethaler P (2006) An ALOHA protocol for multihop mobile wireless networks. *IEEE Trans Inf Theory* 52(2): 421–436
- Jindal N, Andrews JG, Weber S (2008a) Bandwidth partitioning in decentralized wireless networks. *IEEE Trans Wirel Commun* 7(12):5408–5419
- Jindal N, Weber S, Andrews JG (2008b) Fractional power control for decentralized wireless networks. *IEEE Trans Wirel Commun* 7(12):5482–5492
- Jindal N, Andrews JG, Weber S (2011) Multi-antenna communication in ad hoc networks: achieving MIMO

- gains with SIMO transmission. *IEEE Trans Commun* 59(2):529–540
- Kalamkar S, Haenggi M (2018) The spatial outage capacity of wireless networks. *IEEE Trans Wirel Commun* 17(6):3709–3722
- Malanchini I, Weber S, Cesana M (2015) Stochastic characterization of the spectrum sharing game in ad-hoc networks. *Elsevier Comput Netw* 81:63–78
- Vaze R, Truong KT, Weber S, Heath RW (2011) Two-way transmission capacity of wireless ad-hoc networks. *IEEE Trans Wirel Commun* 10(6):1966–1975
- Weber S, Andrews JG (2012) Transmission capacity of wireless networks. *Found Trends® Netw* 5(2–3): 109–281
- Weber S, Yang X, Andrews JG, de Veciana G (2005) Transmission capacity of wireless ad hoc networks with outage constraints. *IEEE Trans Inf Theory* 51(12):4091–4102
- Weber S, Andrews JG, Jindal N (2007a) The effect of fading, channel inversion, and threshold scheduling on ad hoc networks. *IEEE Trans Inf Theory* 53(11):4127–4149
- Weber S, Andrews JG, Yang X, de Veciana G (2007b) Transmission capacity of wireless ad hoc networks with successive interference cancellation. *IEEE Trans Inf Theory* 53(8):2799–2814
- Weber S, Andrews JG, Jindal N (2010) An overview of the transmission capacity of wireless networks. *IEEE Trans Commun* 58(12):3593–3604
- Wildman J, Weber S (2018) On protocol and physical interference models in Poisson wireless networks. *IEEE Trans Wirel Commun* 17(2):808–821
- Wildman J, Nardelli PH, Latva-aho M, Weber S (2014) On the joint impact of beamwidth and orientation error on throughput in directional wireless Poisson networks. *IEEE Trans Wirel Commun* 13(12):7072–7085

Transmission Power Classification

- [Spectrum Sensing with Power Recognition](#)

Transmitting, Passing Forwarding

- [Wireless Sensor Network Security](#)

Trust-Based Routing in Software-Defined Vehicular Ad Hoc Networks

Dajun Zhang

Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

[Trust-based routing in vehicular ad hoc networks;](#)
[Trust-based software-defined routing protocol](#)

Definitions

Vehicular ad hoc networks (VANETs) have turned into a critical technology in smart transportation systems. Because of the high mobility of nodes, VANETs are vulnerable to security attacks. We propose a novel framework of software-defined VANETs with trust management. Specifically, we separate the forwarding plane in VANETs from the control plane, which is responsible for the control functionality, such as routing protocols and trust management in VANETs.

Foundation

In recent years, with the miniaturization of mobile end devices, mobile ad hoc networks (MANETs) become popular in a wide range of fields. For example, they can be used in military, catastrophes, expedition, and so on. A vehicular ad hoc network (VANET) is a type of MANETs in the vehicular environment. With the rising demand of convenient, safe, and efficient transportation, VANETs act as a vital role in intelligent transportation systems (Wei et al. 2015, 2014b; Yu 2016). Network topologies of VANETs always change due to the high mobility of nodes. Meanwhile, VANETs are easy to be attacked by DoS, black hole, and other attacks

(Tyagi and Dembla 2014). So mitigating these attacks is necessary to improve the security of VANETs, and researchers have proposed many security mechanisms in order to enhance the security of VANETs (Wei et al. 2014a, 2015; Wang et al. 2013; Zheng et al. 2012).

Introduction

AODV is one of the most frequently utilized routing protocols in VANETs (Perkins et al. 2003). The main difference between AODV and other VANET routing protocols is that AODV introduces “sequence number” concept, which is utilized to avoid the *count to infinity problem* and to prevent routing loop (Zhang and Gulliver 2005). Specifically, each node in AODV must maintain its own routing table that includes routing information about its neighbor nodes. The operating procedure of AODV can be divided into two main operations: route discovery and route maintenance (Cao and Lu 2010).

Software-defined networking is an emerging network architecture where network control is decoupled from forwarding plane, and it can be directly programmable. Because of the decoupling mechanism of the control and forwarding plane, network nodes only act as forwarding devices. Meanwhile, network control logic is moved into a logic control layer or a networking operating system (Kreutz et al. 2015).

AODV Protocol with Trust-Based Mechanism (TAODV)

In this section, we describe our trust model. We assume that each node in our TAODV broadcasts packets to its neighbors periodically, and the neighbors receive the packets correctly. However, if a node broadcasts multiple packets at the same time, its neighbors only can receive a part of the packets because of some unexpected causes (such as heavy traffic) and malicious attacks (such as black-hole attack). We use a novel concept

forwarding ratio and node trust calculation process (Li et al. 2010) to evaluate our node trust value.

Definition 1(Forwarding ratio): Forwarding ratio is the number of packets received correctly divided by the number of packets forwarded. For example, we assume that a node *a* sends 120 packets to its neighbor node *b*, and node *b* only receives 100 packets because of the packet loss. Meanwhile, node *b* only can forward 80 packets because of its transceiver capability, so the forwarding ratio of node *b* to node *a* is 0.8.

In TAODV, the packets can be divided into two groups: control packets (RREQ, RREP, and RRER) and data packets. The control packets (RREQ and RREP) determine the data transfer path, and forwarding ratio of control packets is an important factor to determine node trust value. Particularly, we assume that the control packet forwarding ratio decides the overall node trust value.

Path Trust Calculation Process

In the route discovery process, when a control packet such as RREQ arrives at a destination node, the routing path from source to destination is computed according to the node trust defined by Sect. 3.1. According to the axiom (Sun et al. 2006), concatenation propagation of trust does not increase trust; the reverse and forwarding path trust value should not be more than the trust value of intermediate nodes. Meanwhile, since the control packet is a crucial factor to determine the node trust value, we add a new field called PacketTrust (PT) into the RREQ and RREP packet format. The path trust value is calculated by the multiplication between two adjacent nodes the packet trust (PT) calculation process can be found in the section of [AODV protocol with Trust based Mechanism](#).

In our TAODV mechanism, there are two main factors influencing the whole network performance. One factor is hop count, and another is path trust value. Our goal is to evaluate the network performance in three different scenarios:

the first one we only consider the path trust factor, the second one we consider both the hop count factor and path trust factor, and the third one we only consider the hop count factor.

Route Discovery Process of TAODV

The traditional AODV protocol aims to select a minimum hop count path to transfer the data packets. By contrast, in our TAODV protocol, we propose a trust-based RREQ (T-RREQ) packet format, which contains the following fields:

(*RREQID*, *HopCount*, *SourceAddr*, *SourceSeq*, *DestAddr*, *DestSeq*, *PacketTrust*)

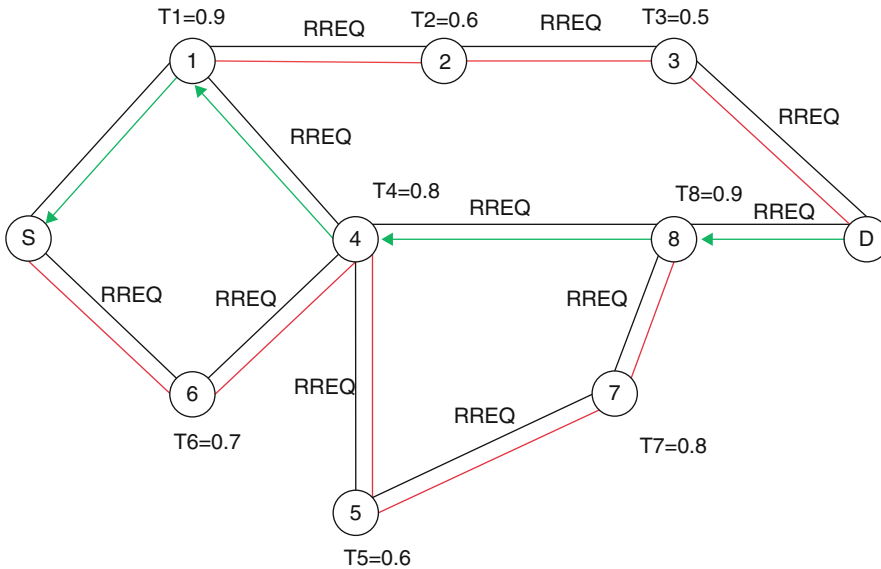
Figure 1 shows an example of reverse path establishment process of TAODV. We assume that the source node needs to initiate the route discovery process. The source first broadcasts T-RREQ packets to its neighbor node 1 and node 6. Meanwhile, the *PacketTrust* field in the T-RREQ is set to 1. The T-RREQ packets arrive at node 1 and node 6, and the path trust is calculated in (4.3). The path trust value from source to node 1 is $T_{s1} = 0.9 \times 1 = 0.9$. The path trust from source to node 6 is $T_{s6} = 0.7 \times 1 = 0.7$. When node 1 and node 6 receive the T-RREQ packets, the

value of *PacketTrust* field of the T-RREQ packets changes to 0.9. Node 4 receives two T-RREQ packets from node 6 and node 1. The routing table of node 4 compares the path trust value. Here the path trust $T_{64} = 0.7 \times 0.8 = 0.56$ and $T_{14} = 0.9 \times 0.8 = 0.72$. So node 4 discards the T-RREQ packet from node 6 because the path trust value from node 6 to node 1 is smaller than the path trust value from node 1 to node 4. After the path selection, node 4 sets up the reverse paths to the source. Similarly, the final reverse path is from destination, via node 7, node 4, and node 1 to the source.

When receiving a T-RREQ, the destination node replies T-RREP back to the source node via the intermediate nodes. Meanwhile, the forwarding paths are established when T-RREP packets pass through the switch nodes. The format of a T-RREP packet contains the following fields:

(*HopCount*, *SourceAddr*, *SourceSeq*, *DestAddr*, *DestSeq*, *PacketTrust*, *lifetime*)

After receiving the RREP packet, the source sends the data packets following the forwarding path that established before to the destination node.



Trust-Based Routing in Software-Defined Vehicular Ad Hoc Networks, Fig. 1 An example of the TAODV calculation process

Software-Defined Vehicular Ad-Hoc Networks Based on TAODV

The framework of SD-TAODV is similar with the traditional SDN architecture. We divide the structure of SD-TAODV into three layers: (1) data forwarding plane operates the TAODV protocol and nodes in the plane supporting OpenFlow protocol; (2) NOS (controller) layer aims to manage the network topology and establishes the data transfer path for the data transmission; (3) application layer controls the forwarding rules, routing tables, and routing protocols. The whole SD-TAODV mechanism virtualizes the VANETs and provides the services for the application layer through the OpenFlow interfaces.

Briefly, if a switch node receives a TAODV control packet (T-RREQ or T-RREP), it sends the packet to the controller to handle. If a switch node receives a data packet, it forwards the packet to its neighbor node(s). The centralized control mechanism in SD-TAODV manages the whole network in the control node. So the control node first needs to know the whole network topology.

The method of discovering the network topology is that the control node sends topology request messages to its neighbors. The topology request message includes the following fields:

(*PacketID*, *ControllerAddr*, *NodeTrustList*, *TopologyList*)

When receiving an OpenFlow message from a forwarding node, the controller determines the type of the message. If the message is *OFPTHello*, the controller responds the message and builds a connection between the node and the controller. If the message is the *OFPTPacketIN*, the control node resolves the message and gets the message information, which includes the details of the T-RREQ and T-RREP packet. As we described before, when a T-RREQ or T-RREP packet arrives at the control node, the controller gets the value in *PacketTrust* field and calculates the path trust value. After finishing the packets handling, the controller sends an *OFPTFlowMod* message to the forwarding node. In SD-TAODV, the forwarding nodes send the *OFPTHello* messages to the control node periodically. If any node in the

network topology receives the response from the control node, this forwarding node builds a connection with the control node. If a forwarding node receives a control packet such as T-RREQ packet from its neighbors, the T-RREQ packet first matches the flow table (we assume that T-RREQ and T-RREP cannot match the flow table). Otherwise, a forwarding node sends this packet to the controller in order to request a new flow table with the *OFPTPacketIn* message. After resolving the packet, the control node responds an *OFPTFlowMod* message back to the node, modifies the flow table, and executes the action set in the flow table to handle this packet.

References

- Cao Z, Lu G (2010) S-AODV: sink routing table over AODV routing protocol for 6LoWPAN. In: Proceedings of the IEEE international conference on networks security wireless communications and trusted computing (NSWCTC), Wuhan, pp 340–343
- Kreutz D, Ramos M, Verissimo P, Rothenberg C, Azodolmolky S, Uhlig S (2015) Software-defined networking: a comprehensive survey. *Proc IEEE* 103(1):14–76
- Li X, Jia Z, Wang L, Wang H (2010) Trust-based on-demand multipath routing in mobile ad hoc networks. *IET Inf Secur* 4(4): 212–232
- Perkins C, Belding-Royer E, Das S (2003) Ad hoc on-demand distance vector (AODV) routing. Technical report, RFC 3561
- Sun Y, Yu W, Han Z, Liu KR (2006) Information theoretic framework of trust modeling and evaluation for ad hoc networks. *IEEE J Sel Areas Commun* 24(2):212–232
- Tyagi P, Dembla D (2014) Investigating the security threats in vehicular ad hoc networks (VANETs): towards security engineering for safer on-road transportation. In: Proceedings of the IEEE international conference on advances in computing, communications and informatics (ICACCI), New Delhi, pp 2084–2090
- Wang Y, Yu FR, Huang M, Chen T (2013) Securing vehicular ad hoc networks with mean field game theory. In: Proceedings of the third ACM international symposium on design and analysis of intelligent vehicular networks and applications (DIVANet'13), pp 55–60. <https://doi.org/10.1145/2512921.2516962>
- Wei Z, Tang H, Yu FR, Wang M, Mason P (2014a) Security enhancements for mobile ad hoc networks with trust management using uncertain reasoning. *IEEE Trans Veh Tech* 63(9):4647–4658

- Wei Z, Yu FR, Boukerche A (2014b) Trust based security enhancements for vehicular ad hoc networks. In: Proceedings of the fourth ACM international symposium on design and analysis of intelligent vehicular networks and applications (DIVANet'14), pp 103–109. <https://doi.org/10.1145/2512921.2516962>
- Wei Z, Yu FR, Boukerche A (2015) Cooperative spectrum sensing with trust assistance for cognitive radio vehicular ad hoc networks. In: Proceedings of the fifth ACM international symposium on design and analysis of intelligent vehicular networks and applications (DIVANet'15), pp 27–33. <https://doi.org/10.1145/2512921.2516962>
- Yu FR (2016) Connected vehicles for intelligent transportation systems [guest editorial]. *IEEE Trans Veh Technol* 65(6):3843–3844
- Zhang Y, Gulliver T (2005) Quality of service for ad hoc on-demand distance vector routing. In: Proceedings of the IEEE international conference on wireless and mobile computing, networking and communications (WiMob), Montreal, pp 192–196
- Zheng D, Yu FR, Boukerche A (2012) Security and quality of service (QoS) co-design using game theory in cooperative wireless ad hoc networks. In: Proceedings of the second ACM international symposium design and analysis of intelligent vehicular networks and applications (DIVANet'12), pp 139–146. <https://doi.org/10.1145/2512921.2516962>

Trust-Based Routing in Vehicular Ad Hoc Networks

- [Trust-Based Routing in Software-Defined Vehicular Ad Hoc Networks](#)

Trust-Based Software-Defined Routing Protocol

- [Trust-Based Routing in Software-Defined Vehicular Ad Hoc Networks](#)

Trusted Computing

- [Verifiable Cloud Computing](#)

Trustworthiness Evaluation

Chen Wang

School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China

Synonyms

[Trustworthiness evaluation](#); [User attribute](#); [Wireless network](#)

Definitions

Trustworthiness evaluation is the act of scoring or grading users by analyzing its attributes. The wireless network systems can decide whether to cooperate with a user by conducting a trustworthiness evaluation of it. How to combine cloud computing technology to securely and efficiently evaluate the trustworthiness of wireless network members while protecting the privacy of each member is one of the focuses of this theme.

Historical Background

User trustworthiness is an important issue that cannot be ignored in network security and information security. The concept of trust for the Internet is first proposed in Grandison and Sloman (2000). The main discussion of Grandison et al. is reputation-based trust. This kind of trustworthiness evaluation may be conducted according to a user's credit card details, purchases, property situation, etc. With the growing development of cashless payments, reputation-based trustworthiness evaluation has demonstrated its role in terms of consumption- and service-level differentiation. In contrast, behavior-based trustworthiness evaluation is proposed in Huang et al. (2006). Huang et al. consider that the trustworthiness of a node in a wireless network represents the accumulative performance of the node in past tasks. Additionally, they also believe that the evaluation of the

trustworthiness can predict the role of the node in future tasks.

Trustworthiness evaluation provides a reference standard for the system's judgment of individuals or interactions between individuals. It is more likely that two mutually trusted individuals will share services and resources. To better develop trustworthiness evaluation, researchers have explored various new evaluation systems and security evaluation schemes in a variety of application environments (Zhang et al. 2013, 2018; Sicari et al. 2015).

Foundations

In fact, both reputation-based and behavior-based trustworthiness evaluation are based on a series of attribute parameters related to the objects. On the basis, the objects are scored or ranked by a certain mathematical method.

First of all, these objects that may be evaluated are referred to herein as users. Both reputation and behavior can be called user attributes. These attributes form the trustworthiness attribute information (TAI). Definition 1 presents the meaning of TAI.

Definition 1 (Trustworthiness attribute information, TAI) TAI represents various attribute parameters of a user which is going to be evaluated by the system. TAI is a collection of parameters obtained by data collectors such as sensors attached to or embedded in users. Set $\mathcal{A} = \{a_1, a_2, a_3, \dots, a_m\}$, where $\mathcal{A}$ represents the set of TAI and $a_1, a_2, a_3, \dots, a_m$ denote different attributes collected by m collectors to reflect the attributes of the user.

As mentioned above, the trustworthiness evaluation may be performed according to a lot of attribute information by some complex computations, which puts high demands on the analyst's computing power and storage capacity. The cloud server is considered as a proper choice to implement the calculations. The information in

the set $\mathcal{A}$ are private data of the evaluated user. The protection of these information is the most important problem to be solved in trustworthiness evaluation.

When the cloud server received the set $\mathcal{A}$, the trustworthiness level (TL) is going to be evaluated according to the attributes in $\mathcal{A}$, which is shown in Definition 2.

Definition 2 (Trustworthiness level, TL) The TL is an evaluation standard provided by the cloud server, according to the TAI uploaded by the user. The evaluation method can be abstracted into a function $\mathfrak{F}$, which is denoted as Eq. (1).

$$TL_{\text{user}} = \mathfrak{F}(\mathcal{A}) \quad (1)$$

It is worth noting that this $\mathfrak{F}$ might be a single function or a series of functions. In addition, $\mathfrak{F}$ also can be a predictive evaluation algorithm constructed by a neural network-based model (Song and Phoha 2004). These can be determined according to the specific requirements of the system.

Applications in Wireless Networks

Trustworthiness evaluation can be widely utilized in wireless networks, especially some mobile networks such as vehicular ad hoc networks (VANETs) (Shen et al. 2017). In these networks, trustworthiness evaluation can be utilized as a credential for admission, as well as a preliminary authentication of the data collected by each individual.

Cross-References

- Access Control
- IoT Security
- Vehicle Ad-Hoc Networks: Services, Security, and Privacy
- Wireless Group Authentication

References

- Grandison T, Sloman M (2000) A survey of trust in internet applications. *IEEE Commun Surv Tutor* 3(4):2–16
- Huang L, Li L, Tan Q (2006) Behavior-based trust in wireless sensor network. In: *Proceedings of Asia-Pacific web conference*. Springer, Berlin/Heidelberg, pp 214–223
- Shen J, Wang C, Castiglione A, Liu D, Esposito C (2017) Trustworthiness evaluation-based routing protocol for incompletely predictable vehicular ad hoc networks. *IEEE Trans Big Data*, <https://doi.org/10.1109/TBDATA.2017.2710347>
- Sicari S, Rizzardi A, Grieco LA, Coen-Porisini A (2015) Security, privacy and trust in internet of things: the road ahead. *Comput Netw* 76:146–164
- Song W, Phoha VV (2004) Neural network-based reputation model in a distributed system. In: *Proceedings of IEEE international conference on e-commerce technology, CEC'2004*. IEEE, Washington, DC, pp 321–324
- Zhang Y, Wang L, Sun W et al (2013) Trust system design optimization in smart grid network infrastructure. *IEEE Trans Smart Grid* 4(1):184–195
- Zhang T, Yan L, Yang Y (2018) Trust evaluation method for clustered wireless sensor networks based on cloud model. *Wirel Netw* 24(3):777–797

Turning Detection in Sandbar Sharks Through Accelerometer Data

Benjamin D. Powell¹, Gang Zhou¹, Daniel P. Crear², Kevin C. Weng², and Wouter Deconinck³

¹Computer Science, College of William and Mary, Williamsburg, VA, USA

²Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, VA, USA

³Physics, College of William and Mary, Williamsburg, VA, USA

Synonyms

Shark turning classification; Shark turning prediction

Definition

Detecting turning using accelerometer data alone is typically impossible, but by putting the indirect “tells” associated with turning through a supervised classifier, periods of frequent turning can be observed.

Historical Background

Marine wildlife is difficult to track or observe through cameras. Instead, attachable tags, typically incorporating accelerometers or gyroscopes, are often used instead. The raw data from those tags is processed and fed into *supervised classifiers*, machine learning algorithms that compare the data to reference data to make predictions. Many researchers (Whitney et al. 2010; Noda et al. 2013; Brownscombe et al. 2014) have successfully used accelerometers to classify long, periodic behaviors in fish, such as swimming, coasting, and mating. However, shorter and more complex behaviors, such as turning, are not as frequently covered.

Existing methods for detecting turning require more electrical power and/or investment than accelerometer tags. Gyroscopes and magnetometers can directly measure orientation (and thus turning), but they require more electrical power than an accelerometer at similar sample rates, limiting tag lifespan. Sonic transmitters, like those used by Holland et al. (1999), can measure turning in certain situations but require a nearby vessel.

Accelerometers are cheap, low-power sensors common in many tags, but detecting turning is difficult, as the yaw – the critical component of turning – cannot be estimated directly. However, measuring secondary behaviors that correlate with turning on a particular animal, in this case sandbar sharks (*Carcharhinus plumbeus*), can be used to predict periods of frequent turning.

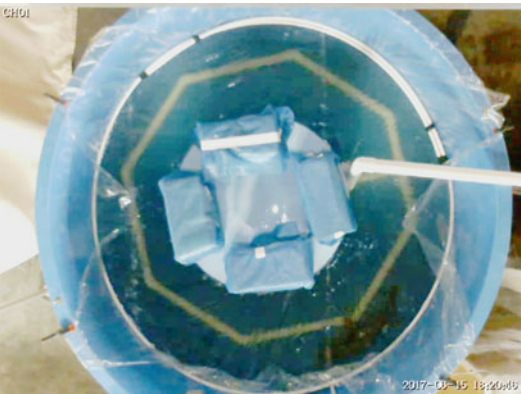
Data Sources and Collection

The data used in this paper consists of 120 h of accelerometer data from four sandbar sharks (*Carcharhinus plumbeus*) and was collected from 2016 to 2017 by the Eastern Shore Lab of the Virginia Institute of Marine Science. As shown in Fig. 1, sharks spent only a small amount of time (3.0%) turning, necessitating heavy preprocessing as covered in the Classification section.

The circular tank, shown in Fig. 2, constricted the movement of the shark, and since the shark is an obligate ram ventilator (Wilga 2009), it must keep swimming in order to breathe. Four behaviors were classified: left/right tight, 180-degree turns and clockwise/counterclockwise swimming around the tank. Ultimately, 11 h and 45 min of data were labeled between all three sharks. One of the authors (BP) exhaustively watched timestamped video data from an overhead camera, marking the time at which a shark ended one behavior and began a new one as well as

Behavior	Frequency
Clockwise	57%
Counterclockwise	40%
Right turn	2.3%
Left turn	0.7%

Turning Detection in Sandbar Sharks Through Accelerometer Data, Fig. 1 Frequency of each behavior in the labeled data



Turning Detection in Sandbar Sharks Through Accelerometer Data, Fig. 2 Still from footage of the tank. (Photo credit: Dan Crear)

the behavior itself. The times for each shark are seen in Fig. 3. For the sake of consistent labeling, a shark was said to start turning (marked as a “left turn” or “right turn”) as soon as its body started to bend in the direction of a turn and was said to start swimming straight (marked as “counterclockwise” or “clockwise”) as soon as its body was fully straightened out. Labels were accurate to the second because of the windowing system discussed in the section below.

Sensors and Feature Selection

A triaxial accelerometer was attached to each shark on the right side of the dorsal fin. Each accelerometer collected data at 25 samples per second and had a maximum range of 16 g and 16 bits of precision.

Classifying shark behavior at a given moment requires knowledge of the data surrounding that moment. Because of this requirement, the data were grouped into 60 sample (2.4 s) windows, 30 of which were shared with the windows around it (15 for each side). Each window was given the label that occupied most of the window. For example, if a window consisted of 20 samples of right turning followed by 40 samples of clockwise swimming, then the window would be classified as clockwise swimming.

The size and overlap were selected by comparing the precision and recall of several options in an ad hoc fashion, though the choices were motivated by Fida et al. (2015), specifically the degree of overlap.

The windows were then converted into *samples*, sets of features fed to the classifier. The features used consisted of several “generic” features included in many papers (Martiskainen et al.

Shark Name	Data Labeled
SB28	02:58:14
SB30	05:43:03
SB34	03:01:05

Turning Detection in Sandbar Sharks Through Accelerometer Data, Fig. 3 Amount of data labeled for each shark

2009; Brownscombe et al. 2014) as well as a novel feature designed specifically for sharks. They are:

- Conditional segmented-mean forward acceleration (described below)
- Mean forward acceleration
- Mean vertical acceleration
- Maximum lateral acceleration
- Mean of squared vertical acceleration $\frac{1}{m} \sum_{i=0}^m a_i^2$
- Mean of squared lateral acceleration

The conditional segmented-mean forward acceleration is a feature that exaggerates the sign of the data in an attempt to distinguish the direction of swimming. First, a threshold t is calculated as the mean forward acceleration of the entire data set. Then, each window X_w is split into the samples above and below t , respectively X_+ and X_- . The final value for window w is

$$\bar{X}'_w = \begin{cases} \bar{X}_+ & \text{if } |X_+| \geq |X_-| \\ \bar{X}_- & \text{otherwise} \end{cases} \quad (1)$$

Originally, this feature was calculated once for every axis (forward, lateral, and vertical), assuming that lateral acceleration would better serve the classifier, as the direction of turning could potentially be derived from centrifugal force. However, the results indicated that forward acceleration had the greatest contribution to the performance of the classifier.

Classification

The samples obtained in Data Sources and Collection were randomly split into a training set (80% of samples) and test set (20% of samples). The training set is used, after some processing, to create the classifier, which then tries to predict what the shark is doing for each sample in the test set. Performance is derived by comparing the predicted and actual behavior for each sample. The small number of turning samples precluded the use of a validation set.

Preprocessing

Before classification, each feature f is standardized. The mean μ_f and standard deviation σ_f of the training set are taken for that feature. Then, for each sample x_f in both the training and test sets, we find $x'_f = \frac{x_f - \mu_f}{\sigma_f}$. This gives each feature the same mean (0) and standard deviation (1).

At this point, the two “minority classes” – left and right turns – make up so little of the dataset that the classifier will naturally choose one of the “majority classes,” clockwise or counterclockwise swimming, every time. To counter this, we artificially resample the data set until there are an equal number of majority and minority samples using a combination of SMOTE (Chawla et al. 2002) and bootstrap sampling (Jacobstein and Porter 2011).

SMOTE is a tool for creating new minority samples from existing ones. The algorithm is fairly simple:

1. For each minority sample a , find the k closest minority samples in feature space.
2. Randomly choose n samples from those neighbors.
3. For each chosen sample b , create a new sample uniformly randomly distributed between a and b .

However, SMOTE does not generate enough minority samples to reach a balance between turning and non-turning samples. Bootstrapping reduces the number of majority samples without reducing the amount of data available to the classifier. To do this, multiple small, balanced “bootstrap samples” are created by randomly sampling from the original dataset. A classifier is trained for each bootstrap sample, creating m bootstrap classifiers. To predict a sample in the test set, each classifier makes its own prediction as a vote, and the class with the most votes is chosen.

In our classifier, 11 bootstrap samples, each with 440 majority samples and 110 minority samples, are created. The majority samples are pulled with replacement, but to reduce the number of duplicates, the minority samples are pulled without replacement. Then, for each bootstrap,

SMOTE is run with $k = 6$ and $n = 3$. Each bootstrap now has the same number of majority and minority samples.

Without any resampling, the classifier was fairly accurate on the whole, successfully classifying 85% of samples. However, it failed to classify any turning samples correctly. Bootstrap sampling alone classified some turning samples but had lower accuracy (60%). The bootstrap + SMOTE technique was more accurate (75%) and had a lower false-positive rate for right turns.

Classifier Design

Each bootstrap classifier was trained with a Support Vector Machine, with an Error-Correcting Output Codes model for multiclassing. SVMs have been used for behavior classification with good results (Martiskainen et al. 2009) but can only distinguish between two classes. ECOC models allow SVMs to distinguish between multiple classes.

Performance, Discussion, and Future Work

The confusion matrix, shown in Fig. 4, shows how the classifier predicted each value in the test set. It reveals that the classifier had a high number of false positives for both left and right turns. Counterclockwise and clockwise swimming have better performance, particularly in distinguishing between one another.

The classifier was good at distinguishing between clockwise and counterclockwise swimming and fairly good at distinguishing between right and left turns. Of the samples correctly classified as turning samples, 86% were correctly predicted to be of a certain direction. The directions of 93% of the samples classified as non-turning were also correctly predicted. Overall, the classifier was about 79% accurate.

The classifier was not, however, able to consistently detect individual left and right turns. The confusion matrix in Fig. 4 shows that a prediction of a turning behavior (either right or left turn) has only a 9% likelihood of actually being a turning

		Predicted			
		L-Turn	R-Turn	CC-Wise	C-Wise
Actual	L-Turn	0	3	7	5
	R-Turn	0	18	11	22
	CC-Wise	10	104	1367	310
	C-Wise	2	100	59	1040

Turning Detection in Sandbar Sharks Through Accelerometer Data, Fig. 4 Confusion matrix for final classifier. The confusion matrix shows the how many samples actually of class i were predicted to be of class j . For example, 22 right turn samples were misclassified as being clockwise

behavior; similarly, 68% of turning samples were classified as non-turning samples.

There remain several possible avenues for performance improvements. A higher ratio of non-turning to turning samples in the bootstraps may have reduced a tendency to misclassify the former samples as the latter. A different majority-minority ratio or degree of SMOTE for each bootstrap may have made the classifier more versatile. Using non-regular windowing (e.g., through an unsupervised novelty detection algorithm) could better match turns to windows. A regression algorithm that outputs the degree and direction of turns could be effective (and useful), perhaps with a gyroscope as a ground truth.

The constricting nature of the tank likely assisted greatly in classifying behaviors. Classifying the behavior of fish in the wild is much more difficult, but it has been done before for non-turning behaviors Noda et al. (2013).

Key Applications

While the classifier is unable to detect individual turns reliably, it could potentially predict periods of rapid turning. The frequency and speed of turns correlates with certain behaviors in many sharks. Holland et al. (1999) found that tiger sharks tend to turn more frequently in shallow water; Kleerekoper et al. (1974) found that nurse sharks turned more frequently when they smelled food downstream. Therefore, it is reasonable to explore similar behavior in sandbar sharks.

The classifier described here is not sufficient for detecting individual turns. However, further efforts may obtain higher performance, and even without it, a much-simplified embedded version of this classifier could activate a tag's gyroscope when it believed it detected multiple turning events in quick succession.

Support

This work was supported in part by the National Science Foundation under grants CNS1253506 (CAREER), DUE-1625872, and PHY-1714792. All vertebrate animal research was conducted in accordance with an approved animal care protocol at the College of William & Mary, IACUC-2017-05-26-12133-kcweng. This work was also supported by the Virginia Institute of Marine Science.

References

- Brownscombe J, Gutoswki L, Danychuk A, Cooke S (2014) Foraging behaviour and activity of a marine benthivorous fish estimated using tri-axial accelerometer biologgers. *Mar Ecol Prog Ser* 505:241–251
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer PW (2002) Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
- Fida B, Bernabucci I, Bibbo D, Conforto S, Schmid M (2015) Varying behavior of different window sizes on the classification of static and dynamic physical activities from a single accelerometer. *Med Eng Phys* 37:705–711
- Holland KN, Wetherbee BM, Lowe CG, Meyer CG (1999) Movements of tiger sharks (*Galeocerdo cuvier*) in coastal Hawaiian waters. *Mar Biol* 134:665–673
- Jacobstein N, Porter B (eds) (2011) 11th IEEE international conference on data mining
- Kleerekoper H, Gruber D, Matis J (1974) Accuracy of localization of a chemical stimulus in flowing and stagnant water by the nurse shark, *Ginglymostoma cirratum*. *J Comp Physiol* 98:257–275
- Martiskainen P, Järvinen M, Skön J-P, Tiirikainen J, Kolehmainen M, Mononen J (2009) Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Appl Anim Behav Sci* 119:32–38
- Noda T, Kawabata Y, Arai N, Mitamura H, Watanabe S (2013) Animal-mounted gyroscope/accelerometer/magnetometer: in situ measurement of the movement performance of fast-start behaviour in fish. *J Exper Mar Biol Ecol* 451:55–68
- Whitney NM, Pratt HLJ, Pratt TC, Carrier JC (2010) Identifying shark mating behaviour using three-dimensional acceleration loggers. *Endanger Species Res* 10:71–82
- Wilga CD (2009) Hyoid and pharyngeal arch function during ventilation and feeding in elasmobranchs: conservation and modification in function. *J Appl Ichthyology* 26:162–166

TV White Space

- [Database-Based Spectrum Access](#)

Two-Phase Flow Microfluidics

- [Communications and Networking in Droplet-Based Microfluidic Systems](#)

Two-Sided Matching

- [Matching for Cooperative Spectrum Sharing](#)

Two-Way

- [Smart Grid Communication Based on IEEE 2030 Standard](#)

Two-Way Relay

- [Fundamental Properties of Wireless Relays and Their Channel Estimation](#)

U

UAV Communication

- [Resource Allocation in UAV-Aided Wireless Networks](#)

UAV-Aided Wireless Network

- [Resource Allocation in UAV-Aided Wireless Networks](#)

Ubiquitous Big Data

Shaohua Wan
School of Information and Safety Engineering,
Zhongnan University of Economics and Law,
Wuhan, Hubei, China

Synonyms

[3V](#); [Intelligent Transportation Systems](#); [Internet of Things](#); [Smart Agriculture](#); [Smart Grids](#); [Smart Health](#); [Ubiquitous big data](#); [Wireless networks](#)

Definition

To date, the definition and characterization of big data remains open. One of representative perceptions made by Gartner is that big data should have 3V feature, i.e., volume, diversity, and timeliness. IDC believes that big data should also have a large value, while IBM holds that it should also increase veracity. Big data analysis is the process of analyzing, comparing, mining, and generating results based on data cleaned, designed and developed according to specific problems, to answer questions or generate new insights (Xiaofeng and Xiang 2013).

Historical Background

The blossom of Internet of Things (IoT) brought about abundant applications, such as LBS (location-based services), traceable logistics, smart home, etc. Though these advanced technologies facilitated people's everyday lives, exponential growth of data rises lots of new challenges, making IoT rely more heavily on big data analyses (Mohammadi et al. 2018; Wan et al. 2019d). In order to effectively exploit the data gathered from IoT, enterprises are obliged to accurately establish a model that can exhume the most valuable information for their

beneficial decision. Drawing support from big data analyses, massive but mussy information can be specifically sanitized to be a significant economic asset, which implies the non-negligible necessity of big data processing in IoT. The data source from wireless networks fall into two categories:

1. Mobile communication data. Nowadays, mobile devices, especially cell phones, constitute a ubiquitous network via ISP (Internet service provider). However, the magnitude and scale of the data collected by infrastructures are heterogeneous and complex, resulting in serious lack of overall integrity. Meanwhile, different applications based on these data are diversely goal-oriented, such as planning a route for a specific user or searching order for a subsequent transaction, inducing distinct requirements including personal information and real-time status as mentioned in the above examples.
2. Sensed data (Wan et al. 2019a). These data are derived from various mechanical and electrical metering devices, such as temperature sensors, accelerometer, cameras, and accessibility detectors or even household appliances. Since a large amount of devices are collecting all kinds of data all the way, IoT provides us affluent information that may help to identify and predict the behavior, such as next steps, subsequent work hours, and possible outcomes of mankind as well as things in terms of well-constructed in-depth data, analyze model, and then automatically formulate instructions to actively torrent potential fault, warning possible disaster or prevent the occurrence of the problem, etc.

Though the concept of big data has been proposed for several years, there are continuous disputes about the definition and characterization of it. Nevertheless, it is believed that the "3V" description presented by Gartner, standing for volume, velocity, and variety, is more accurate. This idea is further expanded as 5V since IDC features big data as provided as great value and IBM paid more attention to their veracity.

As for big data analyses, it can be deemed as the processes of data cleaning, comparison, mining, and results from generation by means of data aggregation, structure unification and developed aiming at specific purposes such as answering questions or producing new insights. The results of the analysis are provided to decision-makers for their decision-maker behavior, no matter they are consistent with the expectation or unpredictable. Considering that the first step of big data analysis is data cleaning, it may induce momentous influence on whole system performance; much attention draws on it. Data cleaning is consisted of three parts, which are data extraction (extract), data conversion (transform), and data loading (load). These steps are generally referred to as data cleaning triad, namely, ETL. As the output of data cleaning, high-dimensional, multimodal big data can be converted into normally structured data which can be further processed more easily. Big data must be combined with the key application to have practical significance.

Key Applications

- **Intelligent Transportation Systems:** Intelligent transportation brings together the big traffic data stream which is abundant and active. From the perspective of the transportation industry, the research on the road traffic flow and a large amount of traffic data is important and significant (van der Heijden et al. 2016; Wang et al. 2016; Wan et al. 2019c; Wan and Goudos 2019). On one hand, dedicated traffic awareness devices deployed on the road surface could provide a large amount of structured and unstructured data, such as induction coil data, video image data, and so on. The big traffic data has the typical characteristics of general big data such as large-scale, real-time, and heterogeneous. It also has the characteristics of the traffic scene itself. The complex nonlinearity and low-value density characteristics of traffic data make it difficult to use big traffic data directly. How to obtain the traffic data quickly and apply the valuable data for traffic control

induction application becomes a fundamental issue. The traditional machine learning and data mining techniques may be not suitable for analyzing the huge data volume of traffic data. The amount of data of urban traffic makes the traditional machine learning technology appear a serious under-fitting state. It is difficult for the traditional data mining method to process such large traffic data types. The complex scenes of urban traffic big data are optimized based on deep learning. Meanwhile, perceptual technology has a positive meaning (Ferdowsi et al. 2017).

- **Smart Grids:** More secure, efficient, and comprehensive data to collect, to communicate, and to process, which enables smart grids to make decisions more accurately and in real time (Wan et al. 2019b; Muralitharan et al. 2018; Gao et al. 2019; Bera et al. 2015; Wang et al. 2019a,b). In smart grids, data keeps being generated in many different devices, which has higher requirements for data acquisition, data transmission, data processing, and real-time visualization. The operation of the power system requires the cooperation of various departments. In particular, the information department needs to not only transmit the data collected from sensors but also analyze and process it. The processing of massive data requires certain technical support. Big data technology is the core technology of data processing in smart grid. Big data analysis provides efficient data processing and computing capabilities to ensure the real time and accuracy of smart grid. The major challenge faced by the smart grid in the era of big data is how to present the vast amounts of different structures and different types of data in a clear, intuitive, and interactive way, which is also the future research direction of big data technology (Erol-Kantarci and Mouftah 2015; Fadel et al. 2015).
- **Smart Health:** The smart medical system uses technologies such as wearable intelligent hardware, intelligent diagnosis, and wireless Internet to integrate premedication, human healthcare, and disease management,

monitoring, diagnosis, and treatment and provide timely feedback (Zhao et al. 2019; Pantelopoulos et al. 2010). The new smart healthcare system will be a huge medical and health system that enables patients to be monitored, diagnosed, and treated anywhere and anytime. It enables users to monitor the physical condition through wearable devices in a convenient environment such as community and home, even in sports and outdoor environments, and transmit data to the cloud platform in real time to monitor physical health through smart medical systems. Moreover, it also makes the patient and the doctor understand and manage the patient's physical condition and disease progression comprehensively. The practice has proved that for multi-class and variability data, the shallow machine learning nonlinear fitting ability is limited. In recent years, big-data-driven deep learning technology has not only achieved certain achievements in the fields of image classification, speech recognition, traffic logistics control, weather forecasting, and driverless driving but also made important breakthroughs in the fields of biological information processing and smart medical treatment.

- **Smart Agriculture:** Agricultural information represented by IoT and intelligent equipment is gradually integrated into the whole process of agricultural production and management, which indicates that agricultural forms and processes have undergone profound changes (Wan and Goudos 2019; Ding et al. 2019; Zhang et al. 2019; Xi et al. 2020; Qi et al. 2020). Firstly, intelligent sensing application has become very popular in smart agriculture. Through the wide application of intelligent sensing devices, the digitalization and perceptibility of agricultural production processes are realized, including crop growth, crop nutrition, livestock growth information, soil parameters, environmental information, climate change, etc. Secondly, comprehensive interconnection and interoperability are integrated in smart agriculture. IoT, sensor network, and Internet are applied in the

agricultural field, which realizes not only the interconnection of farmers, living bodies, and resources and environment but also the interconnection of consumers, agricultural products, and markets. There is more in-depth intelligence. Through advanced technologies such as cloud computing and supercomputers, it could analyze and process the massive perceived data and make agricultural production decisions and agricultural product market management more intelligent. From the perspective of data analysis, it can be summarized into three sentences: one is to generate more data, the other is to transmit more data, and the third is to process more data. It can be seen that the advancement of every technology in the agricultural field has deepened the necessity of the existence and research of big agricultural data to some extent. Big agricultural data refers to the application of big data technology, ideas, and thinking in agriculture. From a deeper level, big agricultural data is a high-level stage of computer technology agricultural application derived from the development of modern information technology that is intelligent, collaborative, precise, networked, and ubiquitous. The semi-structured and unstructured agricultural data can be formulated into the multidimensional, multi-granular, multi-model, multiform massive descriptions, which can be applied to agricultural research, technology promotion, agricultural supply, production management, processing and storage, market sales, and resource environment. It is meant to absorb the value of agricultural data, promote agricultural information consumption, accelerate the transformation, upgrade an agricultural economy, and accelerate the modernization and realization of agriculture. It is a necessary process for agriculture to move to a higher stage.

References

- Bera S, Misra S, Rodrigues JJ (2015) Cloud computing applications for smart grid: a survey. *IEEE Trans Parallel Distrib Syst* 26(5):1477–1494
- Ding S, Qu S, Xi Y, Wan S (2019) Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2019.04.095>
- Erol-Kantarci M, Moutfah HT (2015) Energy-efficient information and communication infrastructures in the smart grid: a survey on interactions and open issues. *IEEE Commun Surv Tutor* 17(1):179–197
- Fadel E, Gungor VC, Nassef L, Akkari N, Malik MA, Almasri S, Akyildiz IF (2015) A survey on wireless sensor networks for smart grid. *Comput Commun* 71:22–33
- Ferdowsi A, Challita U, Saad W (2017) Deep learning for reliable mobile edge analytics in intelligent transportation systems. *arXiv preprint arXiv:1712.04135*
- Gao Z, Xuan HZ, Zhang H, Wan S, Choo KKR (2019) Adaptive fusion and category-level dictionary learning model for multi-view human action recognition. *IEEE Internet Things J* 6(6):9280–9293
- Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M (2018) Deep learning for IoT big data and streaming analytics: a survey. *IEEE Commun Surv Tutor* 20(4):2923–2960
- Muralitharan K, Sakthivel R, Vishnuvarthan R (2018) Neural network based optimization approach for energy demand prediction in smart grid. *Neurocomputing* 273:199–208
- Pantelopoulous A, Bourbakis NG (2010) A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Trans Syst Man Cybern Part C Appl Rev* 40(1):1–2
- Qi L, Zhang X, Li S, Wan S, Wen Y, Gong W (2020) Spatial-temporal data-driven service recommendation with privacy-preservation. *Inf Sci* 515:91–102
- van der Heijden RW, Dietzel S, Leinmüller T, Kargl F (2016) Survey on misbehavior detection in cooperative intelligent transportation systems. *IEEE Commun Surv Tutor* 21(1):779–811
- Wan S, Goudos S (2019) Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput Netw* 168:107036
- Wan S, Gu Z, Ni Q (2019a) Cognitive computing and wireless communications on the edge for healthcare service robots. *Comput Commun*. <https://doi.org/10.1016/j.comcom.2019.10.012>
- Wan S, Li M, Liu G, Wang C (2019b) Recent advances in consensus protocols for blockchain: a survey. *Wirel Netw* 1–15. <https://doi.org/10.1007/s11276-019-02195-0>
- Wan S, Qi L, Xu X, Tong C, Gu Z (2019c) Deep learning models for real-time human activity recognition with smartphones. *Mob Netw Appl* 1–13. <https://doi.org/10.1007/s11036-019-01445-x>
- Wan S, Zhao Y, Wang T, Gu Z, Abbasi QH, Choo KK (2019d) Multi-dimensional data indexing and range query processing via Voronoi diagram for internet of things. *Futur Gener Comput Syst* 91:382–391
- Wang C, Li X, Zhou X, Wang A, Nedjah N (2016) Soft computing in big data intelligent transportation systems. *Appl Soft Comput* 38:1099–1108

- Wang B, Ma H, Wang X, Deng G, Yang Y, Wan S (2019a) Vulnerability assessment method for cyber-physical system considering node heterogeneity. J Supercomput 1–21. <https://doi.org/10.1007/s11227-019-03027-w>
- Wang B, Zhang L, Ma H, Wang H, Wan S (2019b) Parallel LSTM-based regional integrated energy system multienergy source-load information interactive energy prediction. Complexity 7414318
- Xi Y, Zhang Y, Ding S, Wan S (2020) Visual question answering model based on visual relationship detection. Signal Process Image Commun 80:115648
- Xiaofeng M, Xiang C (2013) Big data management: concepts, techniques and challenges. J Comput Res Dev 50(1):146–169
- Zhang R, Xie P, Wang C, Liu G, Wan S (2019) Classifying transportation mode and speed from trajectory data via deep multi-Scale learning. Comput Netw 162:106861
- Zhao Y, Li H, Wan S, Sekuboyina A, Hu X, Tetteh G, Piraud M, Menze B (2019) Knowledge-aided convolutional neural network for small organ segmentation. IEEE J Biomed Health Inform 1363–1373

Ultradense Networks

- [Area Spectral Efficiency of Ultradense Networks](#)

Ultra-reliable and Low-Latency Communications for the Smart Grid

Melike Erol-Kantarci¹ and Antonio Caruso²

¹School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

²Department of Mathematics and Physics 'Ennio De Giorgi, University of Salento, Lecce, Italy

Synonyms

[Low-latency smart grid communications](#); [Ultra-reliable smart grid communications](#)

Definitions

Smart Grids are evolving as the next generation power systems that involve changes in the traditional ways of generation, transmission, and distribution of power. In the new systems, the paradigm of a single flux of energy from few power generator to loads is replaced by multiple flux of distributed power generator and by a multiple flux of data that are needed to observe, analyze, and build the IT infrastructure able to manage and optimize the in-network process.

Ultra-Reliable and Low-Latency Communications (uRLLC) are the key properties of an underlying wireless communication network that supports near real-time operations in the modern smart grid. In addition to smart grid, uRLLC can support a diverse set of recently emerging applications including autonomous vehicles (AVs), tactile applications. A smart grid can use fiber, Ethernet, Power Line Communications (PLC), or wireless mobile networks to support connectivity. The advantage of wireless networks is that they provide ubiquity. Yet, their reliability and latency can be a concern for several smart grid applications. For this reason, uRLLC is a key enabler for a variety of smart grid applications. Existing Long Term Evolution (LTE) networks can barely reach five-nines of reliability under certain circumstances (Baltrunas et al. 2014) while the delay is approximately several tens of milliseconds (Ford et al. 2017). Ultra-reliable and low-latency communications (uRLLC) refer to reliability towards guaranteed five-nines and latency in the order of several milliseconds. This short entry provides the fundamental definitions and key applications of uRLLC along with a discussion of open research issues.

uRLLC and Next Generation Wireless Networks

A recent term used in fifth generation (5G) wireless networks is “*uRLLC*,” an acronym

for *ultra Reliable and Low Latency Communications* along with *eMBB* (enhanced Mobile Broadband) and *mMTC* (massive Machine Type Communications) (Recommendation ITU-R M.2083-0 2015). uRLLC refers to guaranteed five-nines of reliability and several milliseconds of latency. The target use cases include autonomous or self-driving cars and smart grid. On the other hand, eMBB provides greater data-bandwidth and moderate latency improvements targeting augmented, virtual, and mixed reality (AR/VR/MR), as well as Ultra HD or 360-degree streaming video for mobile users. mMTC supports mission critical applications that requires connectivity of a large number of devices.

Smart Grid Requirements

One-way latency is defined as the time it takes for packets to travel over a network from one end of a connection to the other end. Latency can be in the order of seconds, milliseconds, or even microseconds depending on the medium used to transfer the data. For wireless medium, the conventional delays are in the order of several tens of milliseconds.

In the context of smart grid, certain applications are latency-sensitive while some are not, for instance, polling energy consumption data through Advanced Metering Infrastructure (AMI) where smart meter data are usually collected in 10–15 min intervals. On the other hand, real-time demand response, millisecond-level precise load control, control of multiple microgrids, and situational awareness have stringent delay requirements. In particular, most of the control actions in the future smart grids are all about low latency.

Ultra-reliability refers to provisioning of certain level of communication service almost 100% of the time. Situational awareness in smart grid is an example where ultra-reliability is a key issue. Thus, uRLLC is essential for smart grid as most control operations demand tight availability and near real-time service.

Besides, latency requirement can vary from one part of the smart grid to the other, i.e.,

latency requirement for generation, transmission, and distribution system is different from each other. For instance, Supervisory control and data acquisition (SCADA) systems poll the terminals in 2, 4, or 8 s windows and they can tolerate larger latencies and can even make use of satellite links. However, distribution automation with IEC 61850 may require ultra low latencies, in particular for the initial GOOSE messages. Typical delays are in the range of sub-10 ms or less than two cycles of power (60 Hz) at 33.34 ms.

In addition to applications, the architecture of the control systems is important. Historically, all grid management and telemetry systems for the past several decades used centralized control and SCADA. Latency requirement was relaxed compared to the complex multimicrogrids of the future. Such federated grid structures are managed by controllers, and the intelligence is pushed out to the edge of the network and is collocated in the substation managing feeders and perhaps one or two other smaller substations and their associated feeders. Federated grids or multimicrogrids are usually clustered electrical systems that can island from the main grid and self-sustain, if needed. These smaller serving areas may require multiple GOOSE messages sent in many directions for control actions such as tripping switches, connecting or disconnecting renewable energy sources, and monitoring switch conditions.

uRLLC for Community Resilience Microgrid (CRM)

A CRM is a microgrid that is expected to supply uninterrupted electricity during the damage phase and initial recovery period of a resiliency event such as a catastrophic weather conditions. A CRM includes multiple DERs and critical loads that are owned/controlled by different entities within a clearly defined electrical boundary, which are connected via primary distribution lines owned by a local regulated power company (Wu et al. 2017, 2016; Ortmeier et al. 2016). They are also multimicrogrids that can exchange surplus energy among each other.

CRMs exhibit unique structure and bring new challenges for the operation and control, as

well as communication networks. The methods used for control of standalone microgrids, such as droop control, need to be tailored to CRMs. Specifically, as CRMs present a variety of dynamical behaviors ranging from minutes and hours to milliseconds, such as real-time uncertain loads and renewable generation outputs, maintaining reliable operation of CRMs in terms of voltage and frequency stability calls for real-time response and control. In particular, fast response of DERs to instantaneously match critical demands and stabilizing CRM voltage/frequency in real-time are two applications that call for uRLLC (Elsayed and Erol-Kantarci 2018).

uRLLC for Telemetry

Telemetry is one of the fundamental applications required for monitoring the smart grid. Telemetry systems flow in from many devices to the IEC 61850 substation controllers, including telemetry from residential customers and smart meters, reclosers, segmentation switches, power line monitors, transformers, etc. This telemetry provides power quality measurements back to the IEC 61850 substation controller to optimize the voltage and control segmentation and tie switches. uRLLC is expected to support low-latency transmissions of small payloads with very high reliability from a limited set of terminals, which are activated by alarms. Thus, uRLLC is a typical need for telemetry systems.

uRLLC for Situational Awareness

Modern Wide Area Monitoring Systems (WAMS) use Phasor Measurement Units (PMUs) to monitor the power networks in real time. PMUs track system state variables (phasors) for the purpose of state estimation. The sampling rates of PMUs are commonly between 10 and 20 ms (Cosovic et al. 2017). The local systems continuously update its local state estimate based on high-rate PMU inputs, in addition to existing legacy RTU measurements. Considering the distributed nature of state estimation algorithm, along with the need for localized decision making and message exchanging with neighboring units, calls for a network that can support uRLLC.

Open Problems

Heterogeneity

The number of devices over the smart grid is vast. Similarly, a number of communication systems that serve those devices have evolved over time and are operated by multiple utilities. Therefore, the number of available technologies that are required to co-exist is large. Providing low-latency in such heterogenous environment is a fundamental challenge.

Massive Connectivity

Although connectivity of a large number of devices is also studied under mMTC, the uRLLC point of view is concerned about the latency in assigning the available wireless resources to the users or smart grid devices. Clearly, frequency and time resources are limited. Massive number of smart grid devices sharing the same medium with a limited set of resources call for techniques that make the best resource allocation decisions. Thus, further research is needed to solve latency issues in massive connectivity.

Network Slicing

The combination of Software Defined Network (SDN) and Network Function Virtualization (NFV) enables network slicing as a promising way to provide an end-to-end autonomous network. NFV separates network functions from the underlying devices by transforming hardware functions to software which can be installed on commercial off-the-shelf (COTS) equipment. Then, the diverse network functions can be easily deployed in different places to fulfill a specific requirement. Different tenancies can share one physical platform. NFV can reduce the CapEx, for example, globally distributed services of cloud computing can be deployed closer to the users so that infrastructure influence is minimized. With SDN, which provides intelligent network management, the OpEx can be reduced as well. Integration of SDN and NFV offers adaptive traffic steering and high utilization of network resources (Li and Chen 2015). For this purpose, network slicing is emerging as a way of reducing latency that

is incurred at the core of the wireless mobile networks. Some example application areas of network slicing are: intelligent distributed feeder automation, millisecond-level precise load control, information acquirement of low Voltage Distribution Systems and Distributed Power Supplies (Huawei Comm 2015).

Privacy and Security

Smart grid is the producer and the consumer of Big Data. Only the analysis of all the meter readings for a whole year, with smart meters that report the energy consumption every minute, requires terabytes of data to be transmitted, stored, and processed. With the use of cellular wireless networks to transmit data, the security and privacy of the transmitted information emerges as a serious concern (Zhu et al. 2018). The use of uRLLC increases the flexibility and amount of data collected, but at the same time, besides the broadcast nature of the wireless channel is more vulnerable to attacks than wired communication. There are four possible channels for the adversary to target, communications from the smart meter to the substation, from the substation to the cellular tower, between cellular towers, and from the cellular tower to the utility company. The Advanced Metering Infrastructure (AMI) is the first layer that could be exposed to several attacks (Ghosal and Conti 2018) and is also the ones that is more specifically tied to the smart grid concept, so it has been an area of extensive research. Any adversary that is able to monitor the wireless channel could collect large amount of data (even if encrypted) and mine those data to obtain all the messages, or try to recover the cryptographic keys, or exploit regularity in the ciphertext to mine data that reveals the daily schedule of users. Moreover, if the attacker is also able to change control signal from the utility to substations, it could mount critical attacks to the infrastructure. Researchers are now working on how to enforce protection in the smart grid. Typical attacks can be classified in three areas: eavesdropping, traffic analysis, and information theoretic mining. Defense strategies can be employed, for example, using privacy-preserving schemes based on rechargeable-

batteries: the consumption, or the storage of energy in the battery is opaque with respect to the consumption monitored by a smart meters, so it could be used to alter the real consumption and expose a fixed or a random load of a user. The dynamic optimization of protocol parameters of uRLLC to meet the required level of security without decreasing the latency is an active area of research.

Cross-References

- [Advances in Distribution System Monitoring](#)
- [EV Charging Scheduling](#)
- [Mobile Edge Computing: Low Latency and High Reliability](#)
- [Power Line Communications for Grid Discovery and Diagnostics](#)

References

- Baltrunas D, Elmokashfi A, Kvalbein A (2014) Measuring the reliability of mobile broadband networks. In: Proceedings of the 2014 conference on internet measurement conference, ACM, New York, IMC '14, pp 45–58. <https://doi.org/10.1145/2663716.2663725>. <http://doi.acm.org/10.1145/2663716.2663725>
- Cosovic M, Tsitsmelis A, Vukobratovic D, Matamoros J, Anton-Haro C (2017) 5G mobile cellular networks: enabling distributed state estimation for smart grids. *IEEE Commun Mag* 55(10):62–69. <https://doi.org/10.1109/MCOM.2017.1700155>
- Elsayed M, Erol-Kantarci M (2018) Deep Q-Learning for Low-Latency Tactile Applications: Microgrid Communications, accepted to Proc of IEEE SmartGridComm Workshops, Aalborg, Denmark
- Ford R, Zhang M, Mezzavilla M, Dutta S, Rangan S, Zorzi M (2017) Achieving ultra-low latency in 5G millimeter wave cellular networks. *IEEE Commun Mag* 55(3):196–203. <https://doi.org/10.1109/MCOM.2017.1600407CM>
- Ghosal A, Conti M (2018) Key management systems for smart grid advanced metering infrastructure: a survey. ArXiv e-prints 1806.00121
- Huawei Comm SGJ China Telecom (2015) 5G network slicing enabling the smart grid. Document, Huawei Comm, China Telecom, State Grid Jiangsun
- Li Y, Chen M (2015) Software-defined network function virtualization: a survey. *IEEE Access* 3:2542–2553. <https://doi.org/10.1109/ACCESS.2015.2499271>

- Ortmeyer T, Wu L, Li J (2016) Planning and design goals for resilient microgrids. In: 2016 IEEE power energy society innovative smart grid technologies conference (ISGT), pp 1–5, <https://doi.org/10.1109/ISGT.2016.7781248>
- Recommendation ITU-R M.2083-0 (2015) IMT vision – framework and overall objectives of the future development of IMT for 2020 and beyond. Standard, International Telecommunication Union, Geneva, CH
- Wu L, Ortmeyer T, Li J (2016) The community microgrid distribution system of the future. *Electr J* 29(10):16–21. <https://doi.org/10.1016/j.tej.2016.11.008>, <http://www.sciencedirect.com/science/article/pii/S1040619016302123>
- Wu L, Li J, Erol-Kantarci M, Kantarci B (2017) An integrated reconfigurable control and selforganizing communication framework for community resilience microgrids. *Electr J* 30(4):27–34. <https://doi.org/10.1016/j.tej.2017.03.011>, <http://www.sciencedirect.com/science/article/pii/S104061901730060X>, special Issue: Contemporary Strategies for Microgrid Operation & Control
- Zhu L, Li M, Zhang Z, Du X, Guizani M (2018) Big data mining of users' energy consumption patterns in the wireless smart grid. *IEEE Wirel Commun* 25(1):84–89. <https://doi.org/10.1109/MWC.2018.1700157>

Ultra-reliable Smart Grid Communications

- [Ultra-reliable and Low-Latency Communications for the Smart Grid](#)

Universal Identifier-Based Network

Wei Su, Wei Quan, and Hongke Zhang
Beijing Jiaotong University, Beijing, China

Synonyms

[Resolution mapping](#); [Universal Identifier-based Network](#)

Definition

Universal Identifier-based Network (UIN) can feature autonomic identifier-based mappings and take into account the network behaviors, smart service, resource adaption, and component collaboration.

Historical Background

The current Internet is a distributed interconnected system that supports a variety of heterogeneous networks and services. In the technical sense, its achievements can be attributed to the TCP/IP-based Internet architecture, which was successfully designed and has been continuously improved. The history of the Internet can be divided into several stages with significant milestones.

- **Origin of the Internet:** In 1969, the United States (USA) Advanced Research Projects Agency (ARPA) invented a packet switching network, called ARPANET. This technology was completely different from the circuit switching technology used in the traditional telecommunication networks.
- **Three-level Structure of the Internet:** By 1985, the US National Science Foundation (NSF) had built the National Science Foundation Network (NSFNET) based on six large computing centers.
- **Commercialization of the Internet:** Up to 1993, in order to improve the communication quality of the Internet, the NSFNET was gradually divided into several commercial backbones, which were controlled and managed by different Internet Service Providers (ISPs).
- **Application Proliferation and Mobile Services:** In the twenty-first century, the Internet applications changed from traditional web browsing to content distribution and access (Ahlgren et al. 2012). Besides, many emerging network applications have appeared, such as social networks, cloud computing, online games, and so on.

Foundations

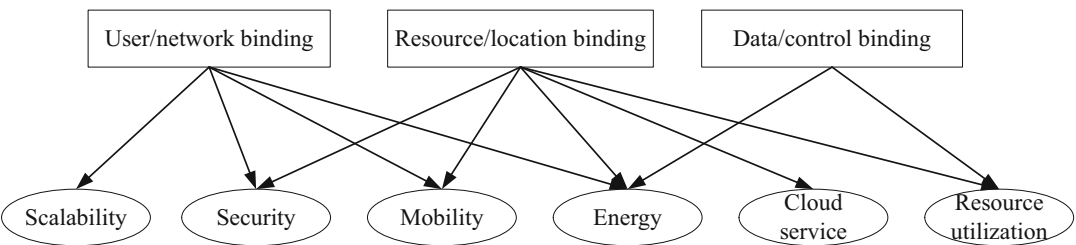
It is explored the root causes of the problems of the current Internet and concluded that it is the triple bindings that result in most of the existing Internet issues. These triple bindings make the Internet relatively static and rigid, which greatly restricts the development of the traditional Internet. The triple bindings refer to the resource/location binding, the user/network binding, and the control/data binding, respectively. Figure 1 illustrates the cause-effect related to the triple bindings and some of the problems in the current Internet.

There are some basic technical terms in UIN. *Core Network (CON)* is in charge of transmitting the backbone data flow. The *CON* is usually operated and managed by the ISP. *Access Network (ACN)* is a network between the *CON* edge interface and the user terminal equipment. *Access Identifier (AID)* is a unique identifier which denotes the identity of the terminal accessed to the network. *Service Identifier (SID)* is an abstract description of service resource information, which is used to uniquely represent a service resource data or a service type. *Routing Identifier (RID)* is an identifier used for the working *CON* equipment. *Connection Identifier (CID)* identifies the process of a user obtaining a service. *Resolution Mapping (RM)* is used to generate the corresponding relationship between different identifiers. *Access Switching Router (ASR)* is located at the edge of an *ACN*. It is responsible for the access of various fixed/mobile terminals and ad-hoc networks. *General Switching Router (GSR)* is the backbone routing equipment in *CON*, which is responsible for the unified routing and forwarding in *CON*. *Identifier*

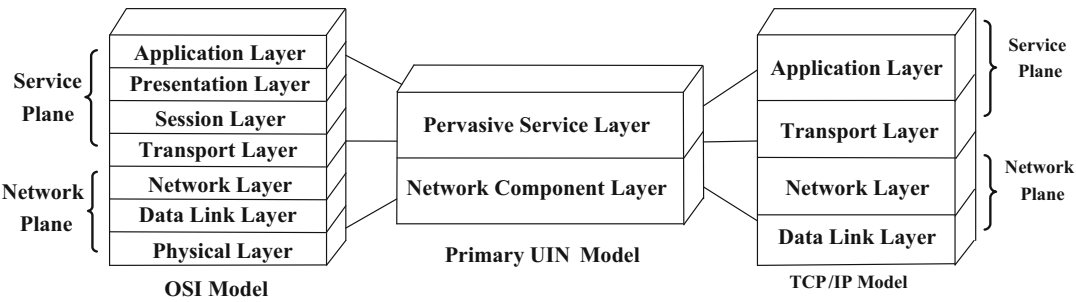
Mapping Server (IDMS) is used to manage and store the mapping rules and algorithms of the UIN. *Authentication Center (AC)* is used to perform the bidirectional authentication process when a node accesses the network.

Different from the traditional telecommunication network, Internet is initially designed for data transmission. It is composed of a series of user access equipment, network switching/routing equipment, and specific service servers. Based on the above analysis, we recognize that the above network architectures share several common features. The first one is that network equipment is interconnected to form a network infrastructure. The second one is that the service requester and the service provider are connected due to the service provision. Therefore, based on the above two features, we originally create the novel UIN architecture which contains only two layers, *Network Component Layer* and *Pervasive Service Layer*, by comparing with the *OSI reference model* (Zimmermann 1980) and the *TCP/IP reference model*. The comparison and relationship of these three network architectures are shown as in Fig. 2.

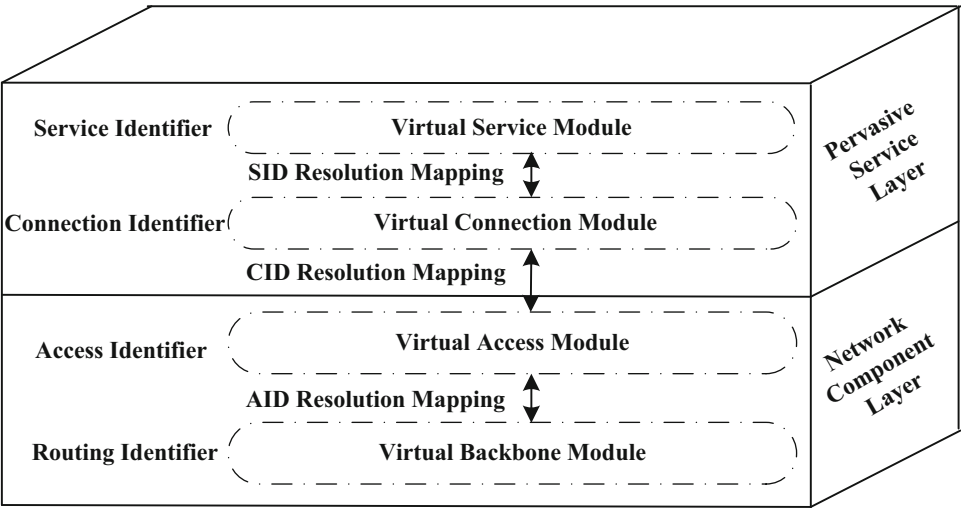
This two layer-based reference model absorbs the characteristic of existing network architectures but removes their disadvantages (Zhang and Luo 2013; Li et al. 2014). It simplifies the design of network and improves the flexibility of network design (Qiu et al. 2013). In UIN, the Network Component Layer brings in the *Virtual Access Module (VAM)* and the *Virtual Backbone Module (VBM)*. The Pervasive Service Layer introduces the *Virtual Service Module (VSM)* and the *Virtual Connection Module (VCM)*. Meanwhile, four kinds of identifiers are introduced, i.e., *SID*, *CID*, *AID*, and *RID*. To bet-



Universal Identifier-Based Network, Fig. 1 The triple bindings and caused problems



Universal Identifier-Based Network, Fig. 2 The comparison of three network architectures



Universal Identifier-Based Network, Fig. 3 Foundational reference model of UIN

ter cover the mentioned requirements, we establish the interaction between the two layers to connect four different function modules as shown in Fig. 3.

All the above modules are responsible for different network functions and building the basic function units of the network communication.

- VAM carries out the AID-related functions, which are responsible for completing unified access of all kinds of network terminals such as the fixed networks, the mobile networks, and the sensor networks. It is also responsible for providing data switching and forwarding in ACN.
- VBM carries out the RID-related functions, which are responsible for switching, routing

and forwarding of data packets, as well as maintaining the interconnection of network equipment.

- VSM carries out the SID-related mechanisms, which are responsible for the unified description, management of all kinds of network service resources, and providing unified service interface for user application\break{} programs.
- VCM carries out the CID-related mechanisms, which are responsible for generating virtual access channel for specific network service according to the service demand and real-time network status (Luo et al. 2013). Meanwhile, it also determines the corresponding network service interface for data received from different connection channels.

Key Applications

Universal Identifier-based Network is a novel network architecture, which can improve the end-to-end performance.

References

- Ahlgren B, Dannewitz C, Imbrenda C, Kutscher D, Ohlman B (2012) A survey of information-centric networking. *IEEE Commun Mag* 50(7):26–36
- Li X, Zhou H, Luo H, Zhang H, Feng Q, You I (2014) HMS: a hierarchical mapping system for the Locator/ID separation network. *Comput Inform* 32(6):1229–1255
- Luo H, Lin Y, Zhang H, Zukerman M (2013) Preventing DDoS attacks by identifier/locator separation. *IEEE Netw* 27(6):60–65
- Qiu F, Luo H, Li X, Xue M, Dong P, Zhang H (2013) A mapping forwarding approach for supporting mobility in networks with identifier/locator separation. *Int J Commun Syst* 26(5):626–643
- Zhang H, Luo H (2013) Fundamental research on theories of smart and cooperative networks. *Acta Electron Sin* 41(7):1249–1254
- Zimmermann H (1980) OSI reference model – the ISO model of architecture for open systems interconnection. *IEEE Trans Commun* 28(4):425–432

Unreachable Electronic Mail

► Mixed Network

Urban Big Data

Hang Shen
Department of Computer Science and
Technology, Nanjing Tech University, Nanjing,
Jiangsu Province, China

Synonyms

City big data

Definition

Urban big data is a massive amount of dynamic and static data generated from the subjects and objects including various urban facilities, organizations, and individuals, which have been being collected and collated by city governments, public institutions, enterprises, and individuals using a new generation information technologies. Big data can be shared, integrated, analyzed, and mined to give people a deeper understanding of the status of urban operations and help them make more informed decisions on urban administration with a more scientific approach, thereby optimizing the allocation of urban resources, reducing the operating costs of the urban system, and promoting the safe, efficient, green, harmonious, and intelligent development of the cities as a whole. In addition to their general features (e.g., volume, velocity, variety, veracity, and value), urban big data also has the additional features below (Pan et al. 2016):

- **Hierarchy:** For example, electronic medical records are categorized by hospital or region, while medical images are categorized in terms of individual medical devices and hospitals. Meanwhile, health data can be categorized in terms of individuals, hospitalized patients, communities, or health and anti-epidemic authorities. The hierarchy of urban big data deeply reflects the organizational hierarchy of a city's physical and social systems.
- **Integrity:** As an urban system evolves, the data coverage of each subsystem becomes increasingly broad. In recent years, for example, the inclusion of environmental protection data in China's cities has improved rapidly. Due to the rapid improvement of data integrity, urban big data has acquired the capacity to uncover the overall dynamics of urban development increasingly accurately.
- **Correlation:** Types of urban data are highly correlated with each other. For example, information about urban logistics is included not only in the data of logistics enterprises but also in the data of the manufacturing, commercial, and transport industries, and even in

the financial industry. Such correlations can be used, not only for mutual corroboration but also for cooperative reasoning and mining rules of cities operation.

Historical Background

The rise in the population, manufacturing, and traffic of its cities is becoming increasingly intense and complex leading to a variety of urban diseases such as rapid population growth, traffic jams, environmental deterioration, housing shortages, employment problems, and public safety challenges. This is just a short list of the side effects of urbanization while there is a host of other less prominent policy problems facing policymakers (Pan et al. 2016). All of these factors have become serious constraints upon the healthy and sustainable development of urban ecosystems (Pan 2016). The advent of “urban big data” has attracted wide attention from academics and industry because it can provide new potential platforms for resolving various “urban diseases.”

Foundations

Urban big data has become an important strategic resource for the development of smart city and strategic direction. As a city evolves toward informatization and intelligence, numerous information bases and data centers have been emerging, which should be properly interconnected to form urban big data. Urban big data can be converged, analyzed, and mined with depth via the Internet of Things, cloud computing, and artificial intelligence technology to help people understand the forces guiding the development for each layer of the urban system to assist the government and society with decision-making and urban planning to achieve the intelligent administration of the city. Meanwhile, urban big data will bring about profound changes to the operating mode of various urban sectors and speed the transformation and upgrading of traditional industries as well as the development of emerging industries (Pan et al. 2016).

Key Applications

Urban big data provides not only a new idea to the in-depth study of urban operations and development (Foth et al. 2011; Wu et al. 2014) but also a new opportunity to renew the competitive advantage of cities (Kitchin 2014). In the context of rapid global informatization, big data has become a vital strategic resource for every city. Strengthening urban competitiveness requires that every city make full use of its advantages in scale, quality, and applications, to tap into and unleash the potential value of data resources, and improve the socioeconomic benefits of big data. Meanwhile, big data has also become a new driving force of urban economic transformation (Lepri et al. 2015). Specifically, big data plays an important role in the following: promoting web-based sharing, intensive integration, and collaborative utilization of production factors; facilitating innovation in the business and circulation modes for production materials, technologies, human resources, and funds; and improving enterprises’ core value and strengths. In addition, the use of big data constantly gives birth to new business patterns and new economic growth points. Last, but not least, big data provides a new way to improve the administrative capacity of governments (Taylor and Richter 2015). Using big data can reveal the latent relationships beyond the reach of traditional technological methods for identifying correlations among seemingly unrelated knowledge and transform such information into new knowledge – and diagnose and evaluate urban development via qualitative and quantitative analysis. Accordingly, big data helps governments improve their data-driven decision-making ability and provide a new means of solving complex social problems (Pan et al. 2016).

Analysis and Processing Techniques

Urban big data is the foundation of a smart city for normal operation, decision-making, control, and display. The ability to intelligently process and analyze urban big data is critical. Obtaining

valuable information from the data and making decisions rely largely on visualization techniques with intelligent algorithms. The basic idea is to take advantage of human's cognitive abilities in visualizing information while utilizing computer's automatic analysis. Further integrating interactive analysis methods and interaction techniques, visual analysis, and processing help people to understand the information, knowledge, and wisdom behind urban big data directly and effectively. On-the-fly data analysis and processing have become increasingly important, and the traditional store-then process approaches have proven to be no longer valid as the capacity of the dataset increases.

Studies have shown that on-the-fly data analysis and processing becomes increasingly important, and the traditional store-then-process approaches have proven to be no longer valid (Singh et al. 2014) as the capacity of the dataset increases.

Cross-References

- [Big Data Networks](#)
- [Smart City](#)

References

- Foth M, Choi JHJ, Satchell C (2011, March) Urban informatics. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. ACM, pp 1–8
- Kitchin R (2014) The real-time city? Big data and smart urbanism. *GeoJournal* 79(1):1–14
- Lepri B, Antonelli F, Pianesi F, Pentland A (2015) Making big data work: smart, sustainable, and safe cities, *EP J Data Science*, pp 16-1–16-4
- Pan Y (2016) China's urban infrastructure challenges. *Engineering* 2(1):29–32
- Pan Y, Tian Y, Liu X, Gu D, Hua G (2016) Urban big data and the development of city intelligence. *Engineering* 2(2):171–178
- Singh D, Tripathi G, Jara AJ (2014) A survey of Internet-of-Things: future vision, architecture, challenges and services. In: *IEEE world forum on Internet of Things (WF-IoT)*, pp 287–292
- Taylor L, Richter C (2015) Big data and urban governance. In: *Geographies of urban governance*. Springer, Cham, pp 175–191
- Wu X, Zhu X, Wu GQ, Ding W (2014) Data mining with big data. *IEEE Trans Knowl Data Eng* 26(1):97–107

User Attribute

- [Trustworthiness Evaluation](#)

User Authentication

- [Radio-Frequency Fingerprinting-Based Physical Layer Identification](#)

User Device Identification

- [Radio-Frequency Fingerprinting-Based Physical Layer Identification](#)

User Pairing

- [Non-orthogonal Multiple Access \(NOMA\)](#)

Valuations

► [Preferences and Utility Functions](#)

Vehicle Ad-Hoc Networks: Services, Security, and Privacy

Jason M. Carter¹, Qian Chen², and Robert A. Bridges¹

¹Cyber and Applied Data Analytics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

²Electrical and Computer Engineering, University of Texas, San Antonio, TX, USA

Synonyms

[Car-to-car mobile ad-hoc network](#); [Intelligent transportation system \(ITS\)](#); [Inter-vehicle communication system](#)

Definitions

A Vehicular Ad-hoc Network (VANET) is a class of spontaneously and dynamically occurring Mobile Ad-hoc Networks (MANETs). MANETs are comprised of mobile nodes that self organize to form a communication network as needed and are independent of infrastructure support

(Zanjireh and Larijani 2015). Vehicles are the mobile nodes in a VANET, although it is expected that non-mobile nodes, e.g., traffic infrastructure, will also participate.

Historical Background

The growing number of Mobile Ad-hoc Networks, e.g., Bluetooth and 802.11, has given birth to a new type of MANET – Vehicular Ad-Hoc Networks (VANETs). Today, Dedicated Short-Range Communications (DSRC) and Cellular Vehicle-to-Everything (C-V2X) are the two main architectures for vehicular network communication. The 75 MHz of spectrum for vehicular communications was allocated by the U.S. Federal Communications Commission (FCC) to support Intelligent Transportation Services over 20 years ago (Zhang and Delgrossi 2012). This is known as 5.9 GHz DSRC or IEEE 802.11p today. DSRC is viewed as being immediately ready for deployment. For instance, the United States has launched a large-scale connected vehicle pilot in Michigan since 2012 (Bezzina and Sayer 2015), and the similar DSRC allocations and on-road testings have been deployed by many countries; in particular, Japan and China began their DSRC testings in 2006 and 2012, respectively (DENSO 2012).

Smartphone cellular wireless standards (i.e., 4G-LTE) now include low-latency, high-throughput, device-to-device communication, making it a candidate for vehicle-to-vehicle

(V2V) communication. The 4G-LTE-enhanced standard uses the transportation safety spectrum to provide V2V, vehicle-to-infrastructure (V2I), and vehicle-to-person (V2P) connectivity. Traditional cellular channels support vehicle-to-network (V2N) communication with connectivity to the Internet. This release chain is the evolutionary path towards using 5th Generation (5G) wireless for the VANET which has the potential to aid autonomous vehicle collaboration.

Early tests of C-V2X in Europe suggested infrastructure-generated messages could be supported, and small vehicle populations could communicate directly. C-V2X-supported safety applications were demonstrated recently at the Intelligent Transportation System (ITS) World Congress 2018 for the first time (ITS-America 2018). Colorado will be the first to deploy C-V2X in the USA (Qualcomm 2018). In addition, 802.11p has also been improved as 802.11px in 2018 (Filippi et al. 2017).

From the deployment point of view, many experts believe that C-V2X is the logical choice for vehicular networks because its infrastructure already exists to support widespread deployment, while others argue that C-V2X has some technical deficiencies carried over from its phone-based lineage. DSRC, on the other hand, was built for vehicular environments where high speeds, Doppler effects, and high bandwidth demands are the norm.

When 4G-LTE routes V2V messages through the cellular network, DSRC outperforms LTE for vehicle safety applications, such as collision avoidance and the electronic traffic sign (Xu et al. 2017). When 4G-LTE's ad-hoc V2V features are used, C-V2X's performance is comparable to DSRC's (Wang et al. 2017; Gonzalez-Martin et al. 2018). Therefore, researchers and developers have suggested DSRC and C-V2X could co-exist, and DSRC standards could be used in C-V2X to support V2V.

Foundations

An ad-hoc network is formed by nodes cooperatively negotiating a channel within designated

spectrum without involving centralized infrastructure for management services. As in other networks, the Open Systems Interconnection (OSI) model provides an architectural foundation for ad-hoc network communication. Standards and node homogeneity are beneficial in large networks; however, standards and technology that support heterogeneous ad-hoc networks are needed to support the trend of interconnecting more consumer devices.

Vehicle populations on the roads and highways evolve in a dynamic and unpredictable way. All vehicle pairs in an ad-hoc population must trust each other's messages when those messages impact safety. Confidential V2V communication, although secure and private, is untenable in such a dynamic environment. Each vehicle would need to transmit copies of a single message encrypted for each recipient. With millions of registered vehicles, the number of potential interacting pairs is tremendous. Broadcasting messages that any vehicle can read is more suitable in the ad-hoc vehicular network. Hence, ad-hoc networks promote an openness that make securing communication more challenging.

Similarly, privacy concerns of the driver will meet challenges imposed by adoption of VANETs. When precise vehicle position data is available in sufficient quantities, drivers' movement patterns can be analyzed. Time and location not only indicate the places they visit but those places may allow others to infer personal conditions, beliefs, and proclivities. When time-sequenced location information is available, drivers' homes are anchor points that can be used to identify other locations the driver visits. Ad-hoc vehicular networks will transport data containing time-stamped vehicle locations to support many services; therefore, driver privacy must be considered.

Benefits of VANETs

Safety: In the USA, due to human errors, approximately 37,461 people die annually in traffic accidents (NHTSA 2017b). Vehicular communication, especially V2X communication

technology, is staged to become integrated with vehicle controls. This new technology alerts drivers to emergent safety critical situations that cannot be detected visually (NHTSA 2017a). Moreover, it can “see” through traffic and around buildings which will reduce up to 80% of non-impaired-driving crashes and potentially decrease 439K to 615K crashes while saving 987 to 1,366 lives each year (NHTSA 2017a).

Mobility: In 2014, US highway users spent 6.9 billion man hours in traffic congestion. Connected vehicle information feeds ITS applications such as advanced mapping services, prioritized traffic signal timings, synchronized highway merging, and automated parking locators. Vehicle connectivity and signal control applications can reduce travel time of emergency vehicles by up to 23% (Chang et al. 2015). The application of cooperative adaptive cruise control and speed harmonization along with V2V and V2I technology can potentially reduce vehicle travel time on freeways by up to 42% (U.S. DOT 2016). Mobility management applications support route discovery and predictive vehicle maintenance.

Cost: Since 1970, highway fuel consumption has increased 86% (U.S. DOT – Office of Highway Policy Information 2014). Vehicles connected to and managed by a unified communication network platform can reduce personal and societal costs of transportation, and the new industry will create new profit opportunities (Mai and Schlesinger 2011). Vehicle platooning may double road capacity (Lioris et al. 2016).

Environmental Protection: “Green” transportation choices such as eco-signal operations can reduce carbon dioxide emissions by 11% and save more than 3.1 billion gallons of fuel every year (U.S. DOT 2016). When surrounding traffic conditions are communicated, individual vehicles can improve their fuel efficiency by 10–20% (Kamal et al. 2013). By 2050, connected autonomous vehicles (CAVs) could reduce passenger vehicle fuel consumption by as much as 44% (U.S. Energy Information Administration 2017).

In order to realize the safety, mobility, cost, and environmental benefits, a major emphasis

on network security and privacy protection are required.

Security

Vehicles are considered as the first personally owned device that relies on real-time network communication to ensure driver, passenger, and pedestrian’s physical safety. For this reason, ad-hoc vehicle networks must provide security services. The revolving door of highly-mobile devices into and out of the network requires a continuous provisioning, revocation, and monitoring mechanism for a large number of security credentials. The security of the network also relies upon end devices and their application security.

Vehicles already depend on intra-network communications (e.g., Controller Area Network, or CAN) allowing subsystems in each vehicle to communicate and coordinate critical functionality. Widespread adoption of VANETs opens a myriad of attack vectors for adversarial manipulation of each vehicle. As a high-profile example, in 2015 an electronic control unit (ECU) in a Jeep Cherokee was compromised over a cellular network in a proof-of-concept research effort. The ECU provided direct access to the CAN bus, which was used to disable the brakes of the moving Jeep (Miller and Valasek 2015). This terrifying use of a cellular network that allowed device-to-device connectivity resulted in a recall of 1.4 million vehicles for software updates (Greenberg 2015a, b).

VANETs have several high-level security and privacy requirements to mitigate this type of compromise:

- All devices entering and remaining in the system must be certified, trusted, and function correctly.
- Vehicle and infrastructure messages must be verifiable and faithfully represent vehicle state and sender intentions. Safety-related services cannot be provided without trusted communication.

- VANETs must support near real-time communication. Network performance also affects safety.
- Unauthorized information disclosure should be prevented.
- Vehicle owners' privacy, such as illegitimate access to their current or past location, must be prevented unless permission is given (Beresford and Stajano 2003).
- Non-repudiation should be supported. Each message must be traceable by an authority to the sender without breaching a person's right to location privacy.

Cryptosystems form the security foundation and provide message confidentiality, integrity, and authentication in the ad-hoc network. Most encryption strategies use asymmetric key cryptography and manage key distribution using one or more Public-Key Infrastructures (PKI). In a PKI, one or more private and a public key pairs are used for security operations. Public keys have defined validity periods and are signed by authorities to generate certificates releasable to anyone. Message trust is provided through a chain of certificates from an end-entity certificate (e.g., vehicle) to a single root of trust. Verification of digital signatures establishes their trustworthiness.

Elliptic Curve Cryptography (ECC) is well-suited to meet the real-time requirements of vehicular networks, since keys are small relative to their security strength. Fast ECC hardware and software implementations are available and well-tested. Specially protected hardware has been recommended for vehicle-based devices to protect private keys and critical security operations (FIPS 2001; NHTSA 2017a).

Some component (or a set of components) of the PKI must be able to trace each message signature to the signer. A message that does not represent the true state of the sender or is intended to do harm is indicative of misbehavior. Misbehavior may be unintentional, in the case of a malfunctioning vehicle, or it may be intentional, in the case of messages under a compromised certificate. Any form of misbehavior must be detected, reported, adjudicated, and *all the*

certificates issued to a misbehaving device must be identified and communicated to the remaining system participants.

Communication between vehicles and infrastructure (V2I, I2V) can follow traditional network security protocols with one exception: protecting driver location privacy. Infrastructure should be prevented from tracking vehicles unless given permission.

The DSRC security functions specified in the IEEE 1609.2 standard address vehicle impersonation, malicious harm to vehicles, traffic misrouting, signal tampering, and denial-of-service. Effective authentication and non-repudiation can mitigate the majority of these threats. Without authentication, false messages pose a significant safety and operational risk to drivers. In the USA, authentication services specified in IEEE 1609.2 are realized in the Security Credential Management System (SCMS) (Brecht et al. 2018). SAE J2945 provides additional security requirements for vehicle on-board equipment. Non-repudiation and authentication is assured through digital signatures. Even with authentication, denial of service attacks remains a problem. Security services are required to authenticate a portion of the messages they receive and false, but correctly formatted, messages will tie up resources.

In C-V2X, data integrity and confidentiality in V2I and I2V communication is primarily provided through encryption after device authentication. Encryption keys are established using the LTE Authorization and Key Agreement (AKA) protocol. The AKA protocol identifies devices and necessitates interaction with infrastructure; therefore, it is ill-suited for D2D communications. Privacy-preserving, V2V communication over C-V2X will require an alternative authentication mechanism to what is provided today.

Privacy

Today, over 1 billion people use mobility applications to plan routes and avoid traffic (Perez 2016). Vehicle owners, policy makers, and ITS designers recognize the value of real-time location data generated by vehicles and transported

across the network. These applications provide a significant consumer value, but they collect location information from all users in order to provide a service; unfortunately, most drivers fail to consider the larger privacy implications of these applications.

V2V privacy requirements in a vehicular ad-hoc network necessitate authentication without identification. Privacy designs proposed for vehicular ad-hoc networks include short-lived identifiers (pseudonyms), shared identifiers, and privacy-preserving cryptography (Zhang and Delgrossi 2012). In the United States, standards (e.g., SAE J2735 and J2945) currently ensure broadcast messages do not contain persistent identifiers. For example, radio MAC and IP addresses are changed periodically along with security certificates.

When using short-lived pseudonyms, proximate vehicle populations should synchronize pseudonym changes. This is because even small differences in pseudonym change times may aid an adversary in linking a vehicle's pseudonym before and after the change. Short-lived pseudonyms can protect driver's privacy in the following strategies:

1. Vehicles are provisioned batches of unique pseudonym certificates by a PKI authority; each certificate has a short validity period.
2. Vehicles are provisioned batches of shared pseudonym certificates by a PKI authority; each certificate has a short validity period.
3. Vehicles are provisioned a group certificate by a PKI authority; the private key associated with the group certificate allows the vehicle to become its own certificate authority and generate short-lived pseudonym certificates as needed (a hybrid system).

Ideally, a different pseudonym certificate would authenticate each message to better protect driver privacy. However, if each vehicle in the USA drove for 30 min, 4.5 trillion certificates would be needed each day. Generating and provisioning this many certificates is unreasonable, and the amount of computation needed to support security and privacy operations increases

when pseudonyms change. Therefore, the verify-on-demand authentication strategy (Krishnan and Weimerskirch 2011) has been proposed to mitigate such overhead in large vehicle populations. In addition, smaller batches are supplied to each vehicle and vehicles use a pseudonym for a short period of time (e.g., 5 min) can re-use the same pseudonym in a subsequent period. Strategy 2 provisions the same certificate to multiple vehicles. Shared certificates provide exceptional privacy protection; however, revocation of one certificate can have unintended effects on those that share the certificate.

Future Directions

On the dawn of VANET deployment, more changes are expected in the transportation industry in the next 5–10 years than the previous 50 years (Burke 2016; Barra 2016). VANETs promise to improve transportation safety and efficiency, but without careful design VANETs may threaten user security and privacy.

While VANETs are poised for widespread deployment and adoption, fundamental infrastructure for hosting VANETs are in flux, DSRC, and C-V2X being the leading possibilities. VANETs provide real-time situation awareness to vehicles and infrastructure-based services. They will be the foundation of an ITS that will benefit society in four key areas: safety, mobility, cost, and environmental protection. Yet, for the abounding benefits afforded by VANETs, security and privacy vulnerabilities will lead to catastrophic consequences if left unprotected. Developments to build security into VANETs are essential. The privacy ramifications of such a network and the valuable data it transports are in the nascent stages of being studied.

Finally, to reap the promised benefits of VANETs, integrating novel technologies with legacy vehicle systems is a challenge that must be addressed. In particular, the fast-paced evolution of communication technology and the long-lived ownership model of personal vehicles. VANET-enabled technologies must adapt to work both

functionally and securely with industry-adopted, proven systems, such as intra-vehicle CAN networks.

Fifth generation wireless (5G) uses millimeter waves to boost communication bandwidth and decrease latency. It will allow new types of data to be transmitted in near real time. As an example, vehicle perceptions of their environment could be shared over 5G networks. This capability may help facilitate autonomous driving and issue in entirely new ways to get from place to place.

References

- Barra M (2016) The next revolution in the auto industry. <https://www.weforum.org/agenda/2016/01/the-next-revolution-in-the-car-industry/>
- Beresford A, Stajano F (2003) Location privacy in pervasive computing. *IEEE Pervasive Comput* 2(1):46–55
- Bezzina D, Sayer J (2015) Safety pilot model deployment: test conductor team report (report no. DOT HS 812 171). National Highway Traffic Safety, Washington, DC
- Brecht B, Theriault D, Weimerskirch A, Whyte W, Kumar V, Hehn T, Goudy R (2018) A security credential management system for V2X communications. *IEEE Trans Intell Transp Syst* 99:1–22
- Burke K (2016) The auto industry will change more in next five years than prior 50, says GM's president. <https://goo.gl/fw7cem>
- Chang J, Hatcher G, Hicks D, Schneeberger J, Staples B, Sundarajan S, Vasudevan M, Wang P, Wunderlich K (2015) Estimated benefits of connected vehicle applications: dynamic mobility applications, AERIS, V2I safety, and road weather management. United States Department of Transportation. Technical report. Noblis, Washington, DC
- DENSO (2012) Denso to begin vehicle-to-vehicle and vehicle-to-infrastructure technology field tests in China. DENSO News. <https://goo.gl/4YpB4L>
- Filippi A, Moerman K, Martinez V, Turley A, Haran O, Toledano R (2017) IEEE802.11p ahead of LTE-V2V for safety applications. Autotalks NXP
- FIPS P (2001) 140-2. National Institute of Standards and Technology (2001). FIPS PUB 140-2: Security requirements for cryptographic modules 25. Gaithersburg, MD. <https://tinyurl.com/y28rz536>
- Gonzalez-Martin M, Sepulcre M, Molina-Masegosa R, Gozalvez J (2018) Analytical models of the performance of C-V2X mode 4 vehicular communications. arXiv preprint arXiv:180706508
- Greenberg A (2015a) After jeep hack, Chrysler recalls 1.4 m vehicles for bug fix. Google Scholar
- Greenberg A (2015b) Hackers remotely kill a jeep on the highway – with me in it. *Wired* 7:21
- ITS-America (2018) 2018 ITS America annual meeting demonstrations. <https://itsdetroit2018.org/demonstrations/>
- Kamal A, Mukai M, Murata J, Kawabe T (2013) Model predictive control of vehicles on urban roads for improved fuel economy. *IEEE Trans Control Syst Technol* 21(3):831–841. <https://doi.org/10.1109/TCST.2012.2198478>
- Krishnan H, Weimerskirch A (2011) Verify-on-demand-a practical and scalable approach for broadcast authentication in vehicle-to-vehicle communication. *SAE Int J Passenger Cars Mech Syst* 4(2011-01-0584):536–546
- Lioris J, Pedarsani R, Tascikaraoglu FY, Varaiya P (2016) Doubling throughput in urban roads by platooning. *IFAC-PapersOnLine* 49(3):49–54
- Mai A, Schlesinger D (2011) A business case for connecting vehicles executive summary. Technical report. Cisco Internet Business Solutions Group (IBSG)
- Miller C, Valasek C (2015) Remote exploitation of an unaltered passenger vehicle. Black Hat USA 2015
- NHTSA (2017a) Federal motor vehicle safety standards; V2V communications. <https://goo.gl/sEjBSk>
- NHTSA (2017b) USDOT releases 2016 fatal traffic crash data. <https://goo.gl/QFWgFu>
- Perez S (2016) Google maps is turning its over a billion users into editors. <https://techcrunch.com/2016/07/21/google-maps-is-turning-its-over-a-billion-users-into-editors/>
- Qualcomm (2018) Panasonic, Qualcomm and ford join forces on first U.S. deployment for C-V2X vehicle communications in Colorado. PR Newswire. <https://goo.gl/P3Ezdh>
- U.S. DOT (2016) Connected vehicle benefits. Technical report. ITS Joint Program Office. <https://www.its.dot.gov/factsheets/pdf/ConnectedVehicleBenefits.pdf>
- U.S. DOT – Office of Highway Policy Information (2014) Highway finance data collection: motor fuel. <https://www.fhwa.dot.gov/policyinformation/pubs/hf/pl11028/chapter5.cfm>
- U.S. Energy Information Administration (2017) Study of the potential energy consumption impacts of connected and automated vehicles. Technical report. Cisco Internet Business Solutions Group (IBSG). <https://goo.gl/9sXp4m>
- Wang M, Winbork M, Zhang Z, Blasco R, Do H, Sorrentino S, Belleschi M, Zang Y (2017) Comparison of LTE and DSRC-based connectivity for intelligent transportation systems. In: Vehicular technology conference (VTC Spring), 2017 IEEE 85th, IEEE, pp 1–5
- Xu Z, Li X, Zhao X, Zhang MH, Wang Z (2017) DSRC versus 4G-LTE for connected vehicle applications: a study on field experiments of vehicular communication performance. *J Adv Transp* 2017:1
- Zanjireh MM, Larijani H (2015) A survey on centralised and distributed clustering routing algorithms for WSNS. *IEEE Veh Technol Conf* 2015, pp 1–6
- Zhang T, Delgrossi L (2012) Vehicle safety communications: protocols, security, and privacy, vol 103. Wiley, Hoboken

Vehicle Cybersecurity

- [Automotive Security](#)

Vehicle-to-Everything (V2X)

- [Routing for Vehicular Ad Hoc Networks](#)

Vehicle-to-Everything (V2X) Communication

- [Wireless Access in Vehicular Environment](#)

Vehicle-to-Grid

- [Game Theory Meeting Vehicle-to-Grid Regulation: Past, Present, and Future](#)

Vehicle-to-Vehicle Communications

- [Vehicular Ad Hoc Network](#)

Vehicular Ad Hoc Network

Fan Li
Beijing Institute of Technology, Beijing, China

Synonyms

[Inter-vehicle communication](#); [Vehicular ad hoc network](#); [Vehicle-to-vehicle communications](#)

Definitions

Vehicular ad hoc networks (VANETs) are classified as an application of mobile ad hoc network (MANET) that has the potential in improving road safety and in providing travelers comfort (Rasheed et al. 2017). VANETs were first mentioned and introduced (Toh 2002) under “car-to-car ad hoc mobile communication and networking” applications, where networks can be formed and information can be relayed among cars. It was shown that vehicle-to-vehicle and vehicle-to-roadside communication architectures will coexist in VANETs to provide road safety, navigation, and other roadside services. VANETs are a key part of the intelligent transportation system (ITS) framework. The nodes in VANETs collect the information about themselves and the environment around them by the installed variety of sensing equipment, transfer the information to central processing unit by communication technology, and analyze the information by computing technology to obtain the optimal path, report the road timely, and arrange the light cycle.

Historical Background

Vehicular ad hoc network (VANET) typically addresses the issue about alerting drivers about the conditions of roads, traffic, and related aspects that are crucial to safety and to the regulation of vehicle flow. Vehicular ad hoc networks are also called inter-vehicle communications (IVC) or vehicle-to-vehicle (V2V) communications. VANET or IVC has drawn a significant research interests from both academia and industry. One of the earliest studies on IVC was started by JSK (Association of Electronic Technology for Automobile Traffic and Driving) of Japan in the early 1980s. Since 2002, with the rapid development of wireless technologies, the number of papers on VANET or IVC has been dramatically increased in academia, and various new workshops were created to address research issues. On the other

hand, several major automobile manufacturers, such as Audi, BMW, and Fiat, have already begun to invest inter-vehicle (C2c 2006). IEEE also formed the new IEEE 802.11p task group which focuses on providing wireless access for the vehicular environment networks and published the formal 802.11p standard in 2010. VANETs comprise of communication-enabled vehicles which act as mobile nodes as well as routers for other nodes, which has its unique characteristics as follows Li and Wang (2007), Jakubiak and Koucheryavy (2008), Toor et al. (2008), and Al-Sultan et al. (2014):

- Highly dynamic topology and frequently disconnected network. Due to high speed of movement between vehicles, the connectivity of the VANETs could also be changed frequently, and the topology of VANETs is always changing. Especially when the vehicle density is low, it has higher probability that the network is disconnected.
- Sufficient energy and storage. The nodes in VANETs are vehicles, which can be equipped with a sufficient number of sensors and computational resources, such as processors, a large memory capacity, advanced antenna technology, and global position system (GPS). These resources can help nodes to have ample energy and computing power (including both storage and processing).
- Mobility modeling and predication. Due to highly mobile node movement and dynamic topology, mobility model and predication play an important role in network protocol design for VANETs. Moreover, vehicular nodes are usually constrained by prebuilt highways, roads, and streets, so given the speed and the street map, the future position of the vehicle can be predicated.
- Variable network density. The network density in VANET varies depending on the traffic density, which can be very high in the case of a traffic jam, or very low, as in suburban traffic.
- Geographical type of communication. Compared to other networks that use unicast or multicast where the communication end

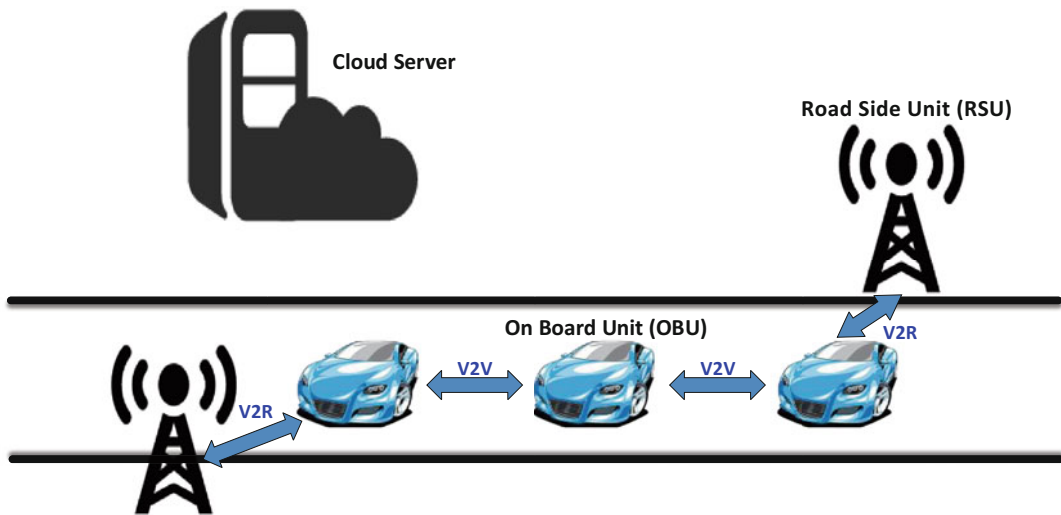
points are defined by ID or group ID, the VANETs often have a new type of communication which addresses geographical areas where packets need to be forwarded (e.g., in safety driving applications).

- Various communication environments. VANETs are usually operated in two typical communication environments. In highway traffic scenarios, the environment is relatively simple and straightforward (e.g., constrained one-dimensional movement), while in city conditions it becomes much more complex. The streets in a city are often separated by buildings, trees, and other obstacles. Therefore, there isn't always a direct line of communications in the direction of intended data communication.

Foundations

MANETs generally do not rely on fixed infrastructure for communication and dissemination of information. VANETs follow the same principle and apply it to the highly dynamic environment of surface transportation. The architecture of VANETs consists of three layers: the end layer, the communication layer, and the cloud layer which is shown in Fig. 1

- The end layer: A variety of smart sensors are mounted on vehicles, which are responsible for acquisition of the vehicle's intelligent information and perception of driving condition and the environment. All these information are gathered in onboard unit (OBU), which is the terminal taking charge of the interior vehicle communication, inter-vehicle communication, and the communication between vehicles and other unit, like cloud server and road side unit (RSU); at the same time, OBU enables the vehicles addressing and having trusted network identification.
- The communication layer: VANETs can be formed by cellular, wireless local area



Vehicular Ad Hoc Network, Fig. 1 VANETs architecture

network (WLAN) and ad hoc way which solve the problem of connectivity between the vehicle and the vehicle (V2V), vehicle and road (V2R), vehicle to infrastructure (V2I), and vehicle to home (V2H). The communication layer realizes vehicle ad hoc network and a variety of communication and roaming between heterogeneous networks ensure real-time performance, serviceability, and network generality on the function and performance. A standard, dedicated short-range communication based on IEEE 802.11p has been established.

- The cloud layer: The cloud layer receives multisource massive amounts of information, which can be audio, video, pictures, and so on, from VANETs. The cloud layer not only receives and stores these information but also analyzes and processes these information and provides a variety of services utilizing these information. These services can be logistics, vehicle repairing, car rental, insurance, emergency rescue, mobile Internet, and so on. This layer is a complex system around the vehicle's data gathering, calculation, scheduling, monitoring, management, and application.

Key Applications

VANETs support a wide range of applications from simple one-hop information dissemination to multi-hop dissemination of messages over vast distances. These applications can be classified according to their primary purpose into safety-related application and client-based application (Toor et al. 2008; Kumar et al. 2013; Martinez et al. 2010):

- Safety-related applications: This category of applications is referred to traveler's safety and management of vehicle traffic on the streets. These applications gather information through the vehicles' sensors in the whole network, process these information, and disseminate the analysis results to all vehicles.
 - Collision avoidance systems lead to the avoidance of many road accidents; these systems are based on the communication between vehicles or between vehicles and infrastructure. This category of applications is designed to send warning messages to vehicles to warn the driver about the current dangerous situation, like the traffic signal, the stop signal, and the intersection.

- Cooperative driving allows vehicles to closely (down to a few inches) follow a leading vehicle by wirelessly receiving acceleration and steering information, thus forming electronically coupled road trains.
- Public safety: where VANET communications, VANET networks, and road safety warning and status information dissemination are used to reduce delays and speed up emergency rescue operations to save the lives of those injured.
- Traffic enhancement: which uses VANET communication to provide up-to-the minute obstacle reports to a vehicle's satellite navigation system.
- Client-based application: These applications provide services with the entertainment (audio, video, game, etc.) and advertising (shops, gas stations, restaurants, etc.). This category of applications is also referred to as non-safety applications and aims to improve drivers and passengers comfort levels (make the journey more pleasant) and enhance traffic efficiency. They can provide drivers or passengers with weather and traffic information and detail the location of the nearest restaurant, petrol station, or hotel and their prices. Passengers can play online games, access the Internet, and send or receive instant messages while the vehicle is connected to the infrastructure network.

Standards

Major standardization of VANET protocol stacks is taking place in the USA, in Europe, and in Japan, corresponding to their dominance in the automotive industry. In the USA, the IEEE 1609 WAVE (Wireless Access in Vehicular Environments) protocol stack builds on IEEE 802.11p WLAN operating on seven reserved channels in the 5.9 GHz frequency band. The WAVE protocol stack is designed to provide multichannel operation (even for vehicles equipped with only a single radio), security, and lightweight application layer protocols. Within the IEEE Communications Society, there is a Technical Subcommittee on Vehicular Networks

and Telematics Applications (VNTA). The charter of this committee is to actively promote technical activities in the field of vehicular networks, V2V, V2R, and V2I communications, standards, communications-enabled road and vehicle safety, real-time traffic monitoring, intersection management technologies, future telematics applications, and ITS-based services.

Protocols

Protocols are a very important issue in VANETs which have been researched by researchers for many years. These protocols all primarily aim to maximize throughput while minimizing packet loss and controlling overhead. One of the major technological challenges faced by VANETs is developing an efficient routing protocol for a highly changeable topology. The current VANET routing protocols can be divided into groups as follow:

- Topological-based routing protocol. This class of routing protocols employs the link information that exists in the network to execute packet forwarding, and the protocols are categorized as (1) proactive (table-driven) routing, which mainly depends on the shortest path algorithms; (2) reactive (on-demand) routing, which is known as on-demand routing protocol because they regularly renew the routing table; and (3) hybrid routing, which are a combination of the reactive and proactive routing protocols to make routing more scalable and efficient.
- Cluster-based routing protocol. Cluster-based routing protocols are more suitable for network cluster topology. Every cluster has one cluster head which is responsible for intra- and inter-cluster management purposes. The intra-cluster nodes interact with one another through direct links, whereas inter-cluster interaction is performed through the cluster headers. The configuration of clusters and the choice of the cluster head are an important issue in cluster-based routing protocols.

- Geocast-based routing protocol. Geocast-based routing is a position-based multicast routing employed to forward a message to all the vehicles in a fixed topographical area. The main goal of this approach is to distribute the packet from the source node to all the other nodes in a particular geographical area or zone of relevance.
- Multicast-based routing protocol. Multicast transmission in the VANET is typically a transmission from a single source to multiple destinations within a specific geographic region and is usually conducted via geocast routing. VANETs are beneficial for multicast protocols because of their wireless nature, which allows a message sent by a node to be broadcast to all nodes within range. VANET nodes do not have to conserve power because vehicles provide substantial power supply for long durations, which enables the nodes to perform complex computational tasks.
- Position-based routing protocol. In position-based routing protocols, all nodes recognize their own locations and their neighbor node geographic locations through position-pointing devices such as GPS. It does not manage any routing table or exchange any information related to the link state with the neighbor nodes. The information from the GPS device is used in making routing decisions. The position-based routing protocols can be classified as non-delay-tolerant network (non-DTN) routing protocols, delay-tolerant network (DTN) routing protocols, and hybrid routing protocols.

Simulations

Prior to the implementation of VANETs on the roads, realistic computer simulations of VANETs using a combination of urban mobility simulation and network simulation are necessary. Typically open-source simulator like SUMO (which handles road traffic simulation) is combined with a network simulator like NetSim (TETCOS), to study the performance of VANETs. A study of cooperative automated driving (Llatser et al. 2017) using Webots and NS3 simulators before

real road test has been performed in the context of the AutoNet 2030 European project.

Cross-References

- [Ad Hoc Networks](#)
- [Dedicated Short-Range Communication \(DSRC\)](#)
- [Delay Tolerant Networks](#)
- [Energy Efficiency of Cellular Networks](#)
- [Evolution of Vehicular Networks in the Transition from 3G to 4G Systems](#)
- [Routing for Vehicular Ad Hoc Networks](#)

References

- Al-Sultan S, Al-Doori MM, Al-Bayatti AH, Zedan H (2014) A comprehensive survey on vehicular ad hoc network. *J Netw Comput Appl* 37(1):380–392
- C2c (2006) Car-to-car communication consortium. <http://www.car-to-car.org>
- Jakubiak J, Koucheryavy Y (2008) State of the Art and research challenges for VANETs, 2008 5th IEEE consumer communications and networking conference, Las Vegas, NV, 2008, pp 912–916
- Kumar V, Mishra S, Chand N (2013) Applications of VANETs: present & future. In: International conference on wireless communications, networking and mobile computing- wicom 2012, Sept 2012, pp 1780–1788
- Li F, Wang Y (2007) Routing in vehicular ad hoc networks: a survey. *IEEE Veh Technol Mag* 2(2):12–22
- Llatser I, Jornod G, Festag A, Mansolino D, Navarro I, Martinoli A (2017) Simulation of cooperative automated driving by bidirectional coupling of vehicle and network simulators. In: Intelligent vehicles symposium, pp 1881–1886
- Martinez FJ, Toh CK, Cano JC, Calafate CT (2010) Emergency services in future intelligent transportation systems based on vehicular communication networks. *IEEE Intell Transp Syst Mag* 2(2):6–20
- Rasheed A, Gilani S, Ajmal S, Qayyum A (2014) Vehicular Ad Hoc Network (VANET): A survey, challenges, and applications. In: Laouiti A, Qayyum A, Mohamad S, Mohamad N (eds) *Vehicular Ad-hoc networks for smart cities: first international workshop*, Springer
- Toh CK (2002) *Ad hoc mobile wireless networks: protocols and systems*. Prentice Hall PTR, Upper Saddle River
- Toor Y, Muhlethaler P, Laouiti A (2008) Vehicle ad hoc networks: applications and related technical issues. *IEEE Commun Surv Tutor* 10(3):74–88

Vehicular Communication System

- [Wireless Access in Vehicular Environment](#)

Vehicular Network Data Offloading

- [Offloading Delay-Tolerable Data Traffic to Connected Vehicle Networks](#)

Vehicular Networks

- [Wireless Sensor Networks Enabled Urban Traffic Management](#)

Vender and Consumer Relationship Management

- [Architecture and Data Management for Smart Community Information Platform](#)

Verifiable Cloud Computing

Cheng Xu, Ce Zhang, and Jianliang Xu
Hong Kong Baptist University, Kowloon Tong,
Hong Kong

Synonyms

[Authenticated query processing](#); [Certified computation](#); [Trusted computing](#)

Definition

Verifiable cloud computing is a way to provide cloud computing services that outsource comput-

ing to untrusted third parties while maintaining the integrity of the computation results.

Historical Background

In cloud computing, the data owner outsources the data storage and query services to a cloud service provider in order to scale up the services with a low cost. However, such an outsourcing model brings about serious issues in computation integrity. As the service provider is not the real owner of the data, it might return incomplete or incorrect results, intentionally or unintentionally. Thus, to ensure computation integrity, the client needs to authenticate the soundness (every result originates from the data owner's database), the completeness (no valid result is missing), and the freshness (the result is up-to-date) of the computation results. To tackle this problem, one early solution is *verification by replication* (Haeberlen et al. 2007). The idea goes as follows. The client outsources the same computing task to multiple workers of different service providers. If not enough results are returned within a reasonable time or the results do not agree, the client would send the task to more other workers. Once a minimum number of the workers agree on the same computation results, the client assumes those results are correct. Although this solution is simple and effective, it assumes failure independence, which however does not hold when facing malicious service providers. In order to provide strong guarantee on the integrity of the computation results against untrusted or even malicious service providers, cryptographically enforced verifiable cloud computing is proposed and studied by a large body of literature. The basic idea is that the service provider should return not only the computation results but also a cryptographic proof, which can be used by the client to establish the soundness, completeness, and freshness of those results.

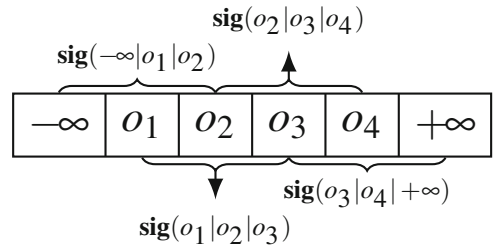
There are several metrics when it comes to evaluate the verifiable cloud computing protocols: (i) preprocessing time, which is the time for the data owner to generate some auxiliary data such as authenticated data structure; (ii) proving

time, which is the time for the service provider to compute the proof; (iii) verification time, which is the time for the client to verify the proof; and (iv) proof size. It is worth noting that the verification time and proof size should ideally be proportional to the size of the computation results and be independent to the size of the whole database. This is particularly important when the client is a mobile user connecting to the cloud through a wireless network.

Foundations

In general, there are two fundamental approaches to support verifiable cloud computing, each with its own advantages and disadvantages. On the one hand, one can design a verifiable scheme specifically based on the computation task. This is often achieved by letting the data owner sign a well-designed *authenticated data structure* (ADS), based on which the service provider can construct corresponding proofs for the outsourced computation. This approach yields low overhead but supports only limited computation tasks. On the other hand, a verification scheme may model the computation task as a general Turing machine. As such, they can support arbitrary tasks at the expense of high and sometimes impractical overhead.

For the ADS-based verification schemes, there are two basic techniques: signature chaining (Fig. 1) and Merkle hash tree (Fig. 2). The former technique uses a public-key cipher with which the data owner can generate digital signatures for each data item using a private key. Consequently, the client can verify the authenticity of the data item using the public key. To establish the completeness of the query results, chaining signatures are generated to capture the correlation of each item and its adjacent items (Pang and Tan 2004). Merkle hash trees (MHTs), on the other hand, are built on index trees (Merkle 1989). Each entry in a leaf node is assigned a digest based on its hashed value, and each entry in an internal node is assigned a digest derived from its child nodes. The data owner signs the root digest of the MHT, which can be used to

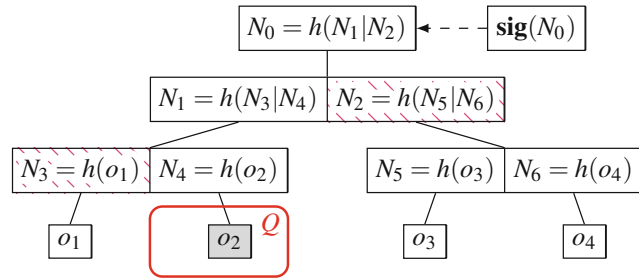


Verifiable Cloud Computing, Fig. 1 Signature chaining

verify any subset of data items. For example, in Fig. 2, a range query Q will return $\{o_2, N_2, N_3\}$, based on which the client can reconstruct the root digest $N_0 = h(h(N_3|h(o_2))|N_2)$ for verification. MHT has been widely adapted to various index structures. Typical examples include the Merkle B-tree for relational data (Li et al. 2006), the Merkle R-tree for spatial data (Yang et al. 2009b; Yiu et al. 2011), the authenticated inverted index for text data (Pang and Mouratidis 2008), the graph metric tree for subgraph similarity search (Peng et al. 2015), and the authenticated prefix tree for multisource data (Chen et al. 2015). It has also been adopted to support authenticated join queries (Yang et al. 2009a), verifiable privacy-preserving location-based services (Hu et al. 2012, 2013; Chen et al. 2013), and queries with fine-grained access control (Xu et al. 2018b). More recently, there have been studies of the ADS schemes on the set-valued data (Papamanthou et al. 2011; Canetti et al. 2014; Zhang et al. 2017b). They utilize a cryptographic set accumulator, which is able to present a (multi-)set with a constant-size digest and support a variety of verifiable (multi-)set operations such as subset, set disjoint, set sum, set intersection, and set union. Based on that, one can achieve verifiable query processing over high-dimensional data (Papadopoulos et al. 2014), verifiable SQL processing (Zhang et al. 2015), or verifiable analytics aggregation queries (Xu et al. 2018a).

In comparison, the general-purpose verifiable cloud computing scheme does not assume any specific properties on the computation task. Instead, the computation task is present as a

Verifiable Cloud Computing, Fig. 2
Merkle hash tree



Boolean or arithmetic circuit, which is Turing complete and thus can be used for arbitrary cloud computation. As the Boolean or arithmetic circuit can be viewed as a series of constraints on the internal computation state as well as the final results, it can be transformed into a so-called quadratic span program. By utilizing certain cryptographic primitives such as pairing, the equivalent program can be verified using a technique known as zk-SNARKs (Parno et al. 2013). In addition to the ability to authenticate arbitrary programs, zk-SNARKs is able to achieve extremely low overhead on verification. The verification time and proof size are both in constant. Further, it leaks no information beyond the computation result, which is crucial for the applications with privacy and confidentiality concerns. However, as a cost of its generality, zk-SNARKs yields high overhead on the preprocessing time and proving time. It is considered impractical to many real-world cloud computing problems. Nevertheless, many studies have been carried out to reduce its proving overhead. Ben-Sasson et al. (2014) propose a zk-SNARK variant that avoids hard-coding the computation program into its verification key and thus reduces the preprocessing cost. Zhang et al. (2017a) propose an interactive protocol for general-purpose SQL queries, whose cost is substantially lower than the original zk-SNARKs scheme. More recently, it has been proposed to model the computation task as a *random access machine* (RAM) program as opposed to a circuit (Braun et al. 2013; Ben-Sasson et al. 2013; Zhang et al. 2018). As a result, it can gain significant performance improvement for some computation tasks. For example, the size of a circuit implementing binary search on a sorted

array is linear in the length of the array, whereas the complexity of a RAM program for binary search is only logarithmic.

Key Applications

Verifiable cloud computing is essential in every cloud computing environment that has security and trust concerns. For example, it is imperative for the application scenario where business intelligence executives make critical, million-dollar decisions such as investing in new businesses based on OLAP queries in the cloud. Similar requirements also exist in many other applications such as scientific research and government policy making.

Cross-References

- [Access Control](#)
- [Authentication](#)
- [Mobile Security and Privacy](#)

References

- Ben-Sasson E, Chiesa A, Genkin D, Tromer E, Virza M (2013) Snarks for c: verifying program executions succinctly and in zero knowledge. In: Canetti R, Garay JA (eds) *Advances in cryptology – CRYPTO 2013*, pp 90–108
- Ben-Sasson E, Chiesa A, Tromer E, Virza M (2014) Succinct non-interactive zero knowledge for a von Neumann architecture. In: *Proceedings of the 23rd USENIX conference on security symposium*, pp 781–796
- Braun B, Feldman AJ, Ren Z, Setty S, Blumberg AJ, Walfish M (2013) Verifying computations with state.

- In: Proceedings of the twenty-fourth ACM symposium on operating systems principles, pp 341–357
- Canetti R, Paneth O, Papadopoulos D, Triandopoulos N (2014) Verifiable set operations over outsourced databases. In: Public-key cryptography – PKC, pp 113–130
- Chen Q, Hu H, Xu J (2013) Authenticating top-k queries in location-based services with confidentiality. *Proc VLDB Endowment* 7(1):49–60
- Chen Q, Hu H, Xu J (2015) Authenticated online data integration services. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp 167–181
- Haerberlen A, Kouznetsov P, Druschel P (2007) Peerreview: practical accountability for distributed systems. *ACM SIGOPS Oper Syst Rev* 41(6):175–188
- Hu H, Xu J, Chen Q, Yang Z (2012) Authenticating location-based services without compromising location privacy. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data, pp 301–312
- Hu H, Chen Q, Xu J (2013) VERDICT: privacy-preserving authentication of range queries in location-based services. In: 2013 IEEE 29th international conference on data engineering, pp 1312–1315
- Li F, Hadjieleftheriou M, Kollios G, Reyzin L (2006) Dynamic authenticated index structures for outsourced databases. In: Proceedings of the 2006 ACM SIGMOD international conference on management of data, pp 121–132
- Merkle RC (1989) A certified digital signature. In: *Advances in cryptology – CRYPTO*, pp 218–238
- Pang H, Mouratidis K (2008) Authenticating the query results of text search engines. In: Proceedings of the VLDB endowment, pp 126–137
- Pang H, Tan KL (2004) Authenticating query results in edge computing. In: Proceedings of the 20th international conference on data engineering, pp 560–571
- Papadopoulos D, Papadopoulos S, Triandopoulos N (2014) Taking authenticated range queries to arbitrary dimensions. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp 819–830
- Papamanthou C, Tamassia R, Triandopoulos N (2011) Optimal verification of operations on dynamic sets. In: *Advances in cryptology – CRYPTO*, pp 91–110
- Parno B, Howell J, Gentry C, Raykova M (2013) Pinocchio: nearly practical verifiable computation. In: 2013 IEEE symposium on security and privacy (SP), pp 238–252
- Peng Y, Fan Z, Choi B, Xu J, Bhowmick SS (2015) Authenticated subgraph similarity search in outsourced graph databases. *IEEE Trans Knowl Data Eng* 27(7):1838–1860
- Xu C, Chen Q, Hu H, Xu J, Hei X (2018a) Authenticating aggregate queries over set-valued data with confidentiality. *IEEE Trans Knowl Data Eng* 30:630–644
- Xu C, Xu J, Hu H, Au MH (2018b) When query authentication meets fine-grained access control: a zero-knowledge approach. In: Proceedings of the 2018 ACM SIGMOD international conference on management of data, pp 147–162
- Yang Y, Papadopoulos S, Kalnis P (2009a) Authenticated join processing in outsourced databases. In: Proceedings of the 2009 ACM SIGMOD international conference on management of data, pp 5–18
- Yang Y, Papadopoulos S, Papadopoulos D, Kollios G (2009b) Authenticated indexing for outsourced spatial databases. *VLDB J* 18(3):631–648
- Yiu ML, Lo E, Yung D (2011) Authentication of moving kNN queries. In: Proceedings of the 27th IEEE international conference on data engineering, pp 565–576
- Zhang Y, Katz J, Papamanthou C (2015) IntegriDB: verifiable SQL for outsourced databases. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp 1480–1491
- Zhang Y, Genkin D, Katz J, Papadopoulos D, Papamanthou C (2017a) vSQL: Verifying arbitrary SQL queries over dynamic outsourced databases. In: IEEE symposium on security and privacy, pp 863–880
- Zhang Y, Katz J, Papamanthou C (2017b) An expressive (zero-knowledge) set accumulator. In: IEEE European symposium on security and privacy (EuroS&P), pp 158–173
- Zhang Y, Genkin D, Katz J, Papadopoulos D, Papamanthou C (2018) vRAM: faster verifiable ram with program-independent preprocessing. In: IEEE symposium on security and privacy

Very Large MIMO

- [Massive MIMO](#)
-

Video Analytics

- [Edge Computing for Real-Time Video Stream Analytics](#)
-

Video Delivery over Wireless Networks

- [Wireless Video Delivery](#)
-

Video Streaming in Wireless Networks

- [Wireless Video Streaming](#)

Video Streaming over Vehicular Networks

Hao Zhou

School of Computer Science and Technology,
University of Science and Technology of China,
Hefei, China

Synonyms

Video streaming over vehicular networks

Definitions

As one of the most important enabling technologies in the envisioned intelligent transportation system (ITS), vehicular networks are designed to provide information exchange via vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications.

Historical Background

With the explosive growth of information technology, vehicular networks contribute to a more efficient driving experience by acting as a promising medium to provide a number of innovation applications, such as traffic monitoring, driving assistance, and multimedia services (Belanović et al. 2010; Ren et al. 2015; Wu et al. 2018).

Vehicular networks employ two different transmission categories, i.e., vehicle-to-infrastructure (V2I) communications, which enable vehicles to communicate with a roadside unit (RSU), and vehicle-to-vehicle (V2V) communications, which enable vehicles to communicate with each other. In 1999, FCC allocated 75 MHz (from 5.850 to 5.925 GHz) bandwidth for dedicated short-range communication (DSRC) in vehicular environments. The IEEE 802.11p/1609 Wireless Access in Vehicular Environment (WAVE) protocol (IEEE Std. 802.11p 2010) and US Department of Telecommunication (US DOT) Federal Highway

Administration DSRC RSU specification document (Chang et al. 2017) have been proposed as the standards for vehicular networks. Meanwhile, the development of cellular communication-supported vehicle-to-everything (V2X) services (3GPP TR 22.885 2015) also received a significant amount of attention from both the industry and academia. Cellular systems, such as Long-Term Evolution (LTE) (Zhou et al. 2015, 2017), are considered as a promising technique for vehicular communication due to their excellent properties, such as high data rate, low latency, large coverage area, high energy efficiency, robust interference control, high penetration rate, and high-speed terminal support (Araniti et al. 2013).

Video streaming over vehicular networks has considerable benefits for both entertainment and road safety. However, high-quality video streaming for fast-moving vehicles encounters fundamental challenges that are attributed to the high mobility and dynamic nature of a network.

Foundations

Channel Estimation in Vehicular Networks

Due to the high mobility of vehicles, the channel condition of each V2I/V2V link may change dramatically during one scheduling period for video streaming. According to Mecklenbrauker et al. (2011), the maximum Doppler frequency in V2V communications can be four times higher than the maximum Doppler frequency in cellular scenarios with the same velocity, which indicates that the time-varying characteristics of vehicular environments are prominent. Therefore, precise channel condition estimation scheme is important and essential for V2I/V2V communications in vehicular networks.

Akhtar et al. perform a realistic analysis of the vehicular network topology characteristics over time and space for a highway scenario (Akhtar et al. 2015). Their investigation results suggest that the most simplistic channel models, including lognormal and unit disk models, fail to provide realistic vehicular network topology characteristics. They further propose a matching

mechanism to tune the parameters of the lognormal model according to the vehicle density and a correlation model to take into account the evolution of the link characteristics over time. The parameters of the proposed model are validated to depend only on the vehicle traffic density based on the real data of four different highways in California.

The substantially predictable vehicle movement makes it possible to estimate the vehicle trajectory in vehicular networks. One can use the position information to obtain partial channel state information (CSI), i.e., the path loss of the channel based on the transceiver distance. The work of Cheng et al. (2015) suggests that resource allocation which utilizes the path loss-based channel state instead of the full CSI is more suitable for vehicular networks; i.e., considering both the path loss and fast fading would not be efficient and effective due to the movement of vehicles.

Video Streaming Over 802.11p-Based Vehicular Networks

Rezende et al. propose a Video Reactive Tracking-based Unicast (VIRTUS) protocol for unicast streaming over vehicular networks (Rezende et al. 2012, 2015). Their research focuses on the suitability of a node to relay packets based on the balance between geographic advancement and link stability. VIRTUS is a reactive solution as its forwarding mechanism is receiver based. It means instead of forwarding nodes choosing the next hop in the route, receiving nodes are responsible for determining if they are supposed to relay the packet. If a node overhears a node closer to the destination forwarding the same packet that it has previously scheduled before its own waiting time expires, it drops the packet.

Asefi et al. propose an adaptive retransmission limit selection scheme to improve the performance of the IEEE 802.11p protocol for video streaming applications over vehicular networks (Asefi et al. 2012). They propose a multi-objective optimization framework to be applied at RSU, which jointly minimizes the probability of playback freezes and start-up

delay of the streamed video at the destination vehicle by tuning the MAC retransmission limit with respect to channel statistics as well as packet transmission rate. Periodic channel state estimation is performed at the RSU using the information derived from the received signal strength (RSS) and Doppler shift effect. Estimates of access probability between the RSU and the destination vehicle are incorporated in the design of the adaptive MAC scheme.

Xing et al. propose an adaptive scheme for scalable video coding (SVC) video streaming services in a highway scenario, where a vehicle can download video data using a direct link or a multihop path to the RSUs via cooperative relay among vehicles (Xing and Cai 2012). Considering the current download speed and the receiver buffer level, the proposed scheme can request an appropriate number of video enhancement layers to improve video quality of experience (QoE).

Belyaev et al. develop and validate a video transmission system for road surveillance applications. They propose low-complexity unequal packet loss protection and rate control algorithms for SVC based on a 3-D discrete wavelet transform (Belyaev et al. 2015). In comparison with a scalable extension of the H.264/AVC standard, the proposed codec is less sensitive to packet losses, has less computational complexity, and provides comparable performance in case of unequal packet loss protection. Measurements obtained in realistic city traffic scenarios demonstrate that good visual quality and continuous playback are possible when the moving vehicle is in the radius of 600 m from the roadside unit.

Video Streaming Over Cellular Communication-Supported Vehicular Networks

Some existing researches focus on video streaming over noncooperative vehicular networks, i.e., V2I communication only. An et al. propose a Resource Allocation and Layer Selection (RALS) algorithm (An et al. 2016) for SVC video streaming scheduling problem, which explicitly takes account of the utility value of each group of picture (GOP) among all the

vehicular users. Xu et al. (2015) introduce a dynamic programming-based resource allocation algorithm for SVC streaming over vehicular networks. Zhou et al. (2016) investigate the resource allocation problem for video multicast services over vehicular networks to answer the questions of which modulation and coding scheme (MCS) should be adopted for each flow in each time segment and how to schedule the radio resources among all the flows. For the multicast service to cover all the recipients, a k-commodity packing-based approximation algorithm is proposed to solve it with low complexity. For the multicast service with adaptive recipients, a heuristic algorithm is proposed to search for proper MCS profile assignments.

As for cooperative vehicular networks, the scheme proposed by Yaacoub et al. is based on grouping moving vehicles into cooperative clusters (Yaacoub et al. 2015). As shown in Fig. 1a, the LTE system is used to send data over long-range cellular links to a selected cluster head, which multicasts the received video over IEEE 802.11p links to vehicles in its cluster. Error concealment techniques are used to compensate the loss of frames that are not delivered on time for real-time video streaming. In addition, resource allocation to select the best subcarriers for LTE transmission is performed in order to enhance the received video quality.

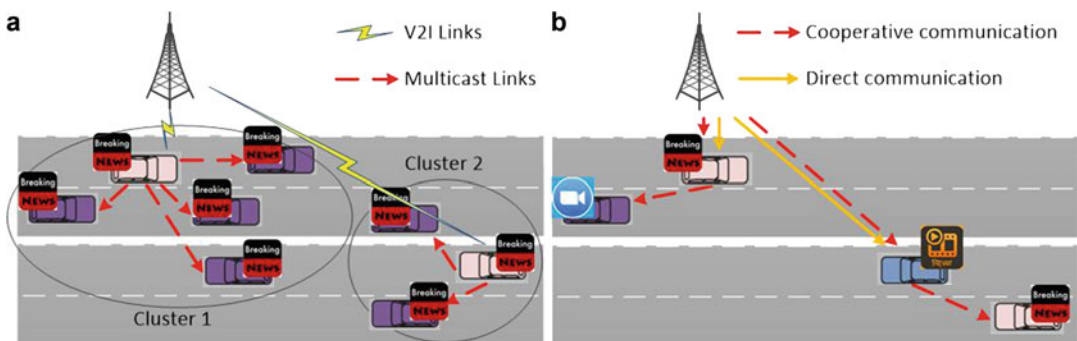
As shown in Fig. 1b, Zhou et al. (2018) consider SVC video streaming over another type of cooperative vehicular networks where the V2I communications are based on LTE links

and the V2V communications occur via another OFDMA-based system using a band of DSRC or a dedicated portion of the LTE band for device-to-device (D2D) communications in overlay mode. They investigate the joint optimization problem about the number of video layers that should be transmitted for each vehicle, the selection of relay vehicles to assist destination vehicles, and the assignment of radio resources to direct and cooperative communications. The optimization problem is decoupled into two subproblems. The lower-level quality of service (QoS)-driven resource allocation subproblem is solved as a generalized maximum concurrent flow problem, and the high-level layer selection subproblem is solved using a tabu search-based metaheuristic algorithm.

Key Applications

Video streaming over vehicular networks is important for both entertainment and road safety services.

For instance, video is the desired media for delivering advertisements and news, which provide passengers with a vast amount of information, more options for entertainment, and higher office efficiency (Zhou et al. 2014; Xing et al. 2016; He et al. 2016). Emergency-related video streaming also has an important role in enhancing road safety (Park et al. 2006; Zhuang et al. 2011).



Video Streaming over Vehicular Networks, Fig. 1 Two different system models for video streaming over cooperative vehicular networks (a) Cluster-based video multicasting. (b) Cooperative-based video unicasting

Future Directions

Vehicular Networks Driven by Deep Learning

With rapid development of artificial intelligence, deep learning has been successfully applied in many areas, including computer vision and speech recognition, and has drawn a lot of attentions of researchers in the vehicular network area (Ye et al. 2018).

In Atallah et al. (2018), the deep reinforcement learning model, namely, deep Q-network, is investigated to learn a scheduling policy from high-dimensional inputs corresponding to the current characteristics of the underlying model. In Sheng et al. (2018), the deep reinforcement learning method has also been used for resource allocation problem in vehicular networks.

However, more studies should be conducted to apply deep learning in vehicular networks. For example, more theoretical analysis should be done to provide guidelines on the architecture design of neural networks and the selection of hyper-parameters.

Cross-References

- [Resource Allocation](#)
- [Vehicular Networks](#)

References

- 3GPP TR 22885 (2015) Technical specification group services and system aspects; Study on LTE support for V2X services, Rel. 14
- Akhtar N, Ergen SC, Ozkasap O (2015) Vehicle mobility and communication channel models for realistic and efficient highway vanet simulation. *IEEE Trans Veh Technol* 64(1):248–262
- An R, Liu Z, Zhou H, Ji Y (2016) Resource allocation and layer selection for scalable video streaming over highway vehicular networks. *IEICE Trans Fundam Electron Commun Comput Sci* 99(11):1909–1917
- Araniti G, Campolo C, Condoluci M, Iera A, Molinaro A (2013) LTE for vehicular networking: a survey. *Commun Mag IEEE* 51(5):148–157
- Asefi M, Mark JW, Shen XS (2012) A mobility-aware and quality-driven retransmission limit adaptation scheme for video streaming over vanets. *IEEE Trans Wirel Commun* 11(5):1817–1827
- Atallah RF, Assi CM, Khabbaz MJ (2018) Scheduling the operation of a connected vehicular network using deep reinforcement learning. *IEEE Trans Intell Transp Syst* 1–14, <http://doi.org/10.1109/TITS.2018.2832219>
- Belanović P, Valerio D, Paier A, Zemen T, Ricciato F, Mecklenbräuker CF (2010) On wireless links for vehicle-to-infrastructure communications. *IEEE Trans Veh Technol* 59(1):269–282
- Belyaev E, Vinel A, Surak A, Gabbouj M, Jonsson M, Egiastian K (2015) Robust vehicle-to-infrastructure video transmission for road surveillance applications. *IEEE Trans Veh Technol* 64(7):2991–3003
- Chang J et al (2017) An overview of US DOT connected vehicle roadside unit research activities. Technical report, United States. Department of Transportation. ITS Joint Program Office
- Cheng X, Yang L, Shen X (2015) D2D for intelligent transportation systems: a feasibility study. *IEEE Trans Intell Transp Syst* 16(4):1784–1793
- He H, Shan H, Huang A, Sun L (2016) Resource allocation for video streaming in heterogeneous cognitive vehicular networks. *IEEE Trans Veh Technol* 65(10):7917–7930
- IEEE Std 80211p (2010) Part 11, Amendment 6: wireless access in vehicular environments (WAVE)
- Mecklenbrauker CF, Molisch AF, Karedal J, Tufvesson F, Paier A, Bernadó L, Zemen T, Klemp O, Czink N (2011) Vehicular channel characterization and its implications for wireless system design and performance. *Proc IEEE* 99(7):1189–1212
- Park JS, Lee U, Oh SY, Gerla M, Lun DS (2006) Emergency related video streaming in vanet using network coding. In: *Proceedings of the 3rd international workshop on vehicular ad hoc networks*. ACM, pp 102–103
- Ren Y, Liu F, Liu Z, Wang C, Ji Y (2015) Power control in D2D-based vehicular communication networks. *IEEE Trans Veh Technol* 64(12):5547–5562
- Rezende C, Ramos HS, Pazzi RW, Boukerche A, Frery AC, Loureiro AA (2012) Virtus: a resilient location-aware video unicast scheme for vehicular networks. In: *2012 IEEE international conference on communications (ICC)*. IEEE, pp 698–702
- Rezende CG, Boukerche A, Ramos HS, Loureiro AA (2015) A reactive and scalable unicast solution for video streaming over vanets. *IEEE Trans Comput* 64(3):614–626
- Sheng Z, Pressas A, Ocheri V, Ali F, Rudd R, Nekovee M (2018) Intelligent 5G vehicular networks: an integration of DSRC and mmwave communications. In: *2018 international conference on information and communication technology convergence (ICTC)*. IEEE, pp 571–576
- Wu C, Liu Z, Zhang D, Yoshinaga T, Ji Y (2018) Spatial intelligence toward trustworthy vehicular IOT. *IEEE Commun Mag* 56(10):22–27

- Xing M, Cai L (2012) Adaptive video streaming with inter-vehicle relay for highway vanet scenario. In: 2012 IEEE international conference on communications (ICC). IEEE, pp 5168–5172
- Xing M, He J, Cai L (2016) Maximum-utility scheduling for multimedia transmission in drive-thru internet. *IEEE Trans Veh Technol* 65(4):2649–2658
- Xu Y, Zhou H, Wang X, Zhao B (2015) Resource allocation for scalable video streaming in highway vanet. In: 2015 international conference on wireless communications & signal processing (WCSP). IEEE, pp 1–5
- Yaacoub E, Filali F, Abu-Dayya A (2015) QOE enhancement of SVC video streaming over vehicular networks using cooperative LTE/802.11 p communications. *IEEE J Sel Top Sign Process* 9(1):37–49
- Ye H, Liang L, Li GY, Kim J, Lu L, Wu M (2018) Machine learning for vehicular networks: recent advances and application examples. *IEEE Veh Technol Mag* 13(2):94–101
- Zhou H, Liu B, Luan TH, Hou F, Gui L, Li Y, Yu Q, Shen XS (2014) Chaincluster: engineering a cooperative content distribution framework for highway vehicular communications. *IEEE Trans Intell Transp Syst* 15(6):2644–2657
- Zhou H, Ji Y, Wang X, Zhao B (2015) ADMM based algorithm for eICIC configuration in heterogeneous cellular networks. In: 2015 IEEE conference on computer communications (INFOCOM). IEEE, pp 343–351
- Zhou H, Wang X, Liu Z, Zhao X, Ji Y, Yamada S (2016) Qos-aware resource allocation for multicast service over vehicular networks. In: 2016 8th international conference on wireless communications & signal processing (WCSP). IEEE, pp 1–5
- Zhou H, Ji Y, Wang X, Yamada S (2017) eICIC configuration algorithm with service scalability in heterogeneous cellular networks. *IEEE/ACM Trans Netw (TON)* 25(1):520–535
- Zhou H, Wang X, Liu Z, Ji Y, Yamada S (2018) Resource allocation for SVC streaming over cooperative vehicular networks. *IEEE Trans Veh Technol* 67(9):7924–7936
- Zhuang Y, Pan J, Luo Y, Cai L (2011) Time and location-critical emergency message dissemination for vehicular ad-hoc networks. *IEEE J Sel Areas Commun* 29(1):187–196

Virtual MIMO

- [Base Station Cooperations Under Imperfect Conditions](#)
- [Network MIMO](#)

Visible Light Communication

- [Indoor Visible Light Communication Networks for Camera-Based Mobile Sensing](#)

Vital Sign Monitoring Using WiFi Signal

- [Sleep Monitoring Using WiFi Signal](#)

Wake-on-Wireless

- ▶ [Radio-on-Demand Wireless LAN](#)

Wake-on-WLAN

- ▶ [Radio-on-Demand Wireless LAN](#)

Wake-Up Control of WLAN

- ▶ [Radio-on-Demand Wireless LAN](#)

WBAN

- ▶ [60 GHz in WBAN and mm-Wave Technologies](#)

Wide Area Monitoring System

- ▶ [Advances in Distribution System Monitoring](#)

Wi-Fi localization

- ▶ [Fingerprinting Localization](#)

Wi-Fi Offloading in Wi-Fi-Cellular Networks

Dongeun Suh¹, Haneul Ko², and Sangheon Pack²

¹Korea University, Seoul, Korea

²Korea University, Seoul, South Korea

Synonyms

[Mobile data offloading techniques in Wi-Fi-cellular networks](#)

Definitions

- *WiFi offloading in Wi-Fi-cellular networks:* WiFi offloading is a promising solution that can cope with the mobile data explosion by migrating traffic from cellular networks to WiFi networks.

Historical Background

For the last several years, the increasing number of wireless devices has lifted mobile data traffic

to a great extent around the world. According to the recent report (Cisco 2017), the compound annual growth rate (CAGR) of the global mobile data traffic is forecasted as 47% from 2016 to 2021. In addition, the global mobile data traffic is expected to grow to 49 exabytes per month by 2021, which is a sevenfold increase over 2016.

To cope with the traffic growth, one of the most straightforward solutions is to increase the cellular network capacity by installing additional base stations (BSs) (e.g., conventional macroBSs and femtoBSs). However increasing the cellular network capacity requires expensive capital expenditure (CAPEX) and operation expenditure (OPEX) (Jung et al. 2014). Also, with the respect to cellular network operators, increasing the capacity is not cost-efficient under a flat price policy where the revenue is independent of data usage.

In this context, leveraging existing radio access technologies (RATs) in heterogeneous networks to offload mobile data traffic can be a practical approach. That is, a mobile node (MN) is capable of accessing various RATs and migrates its traffic from cellular networks to femtocell or WiFi networks. Specifically, since WiFi access points (APs) are widely deployed and they have low CAPEX and OPEX, WiFi offloading is considered as a promising solution to address the mobile data explosion problem in which the mobile data traffic is offloaded through WiFi networks in spatially overlapped WiFi and cellular networks. In fact, WiFi offloading has been widely adopted by a large number of mobile network operators (e.g., Verizon, AT&T, T-Mobile, and Vodafone) (Juniper Research 2013).

Foundations

In general, WiFi offloading techniques can be classified into two types as follows:

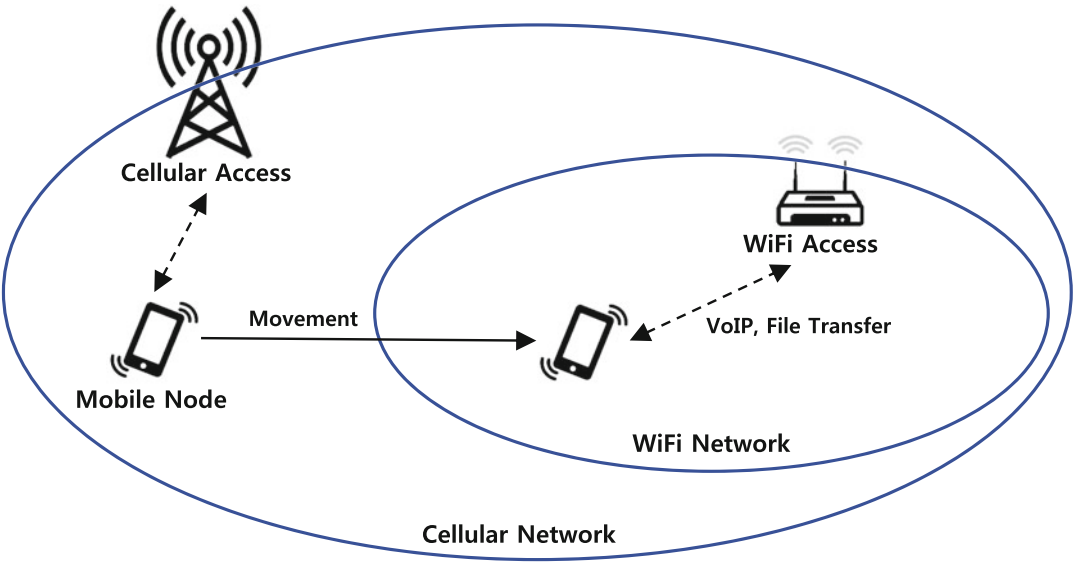
- *On-the-spot WiFi offloading*: In the on-the-spot WiFi offloading technique, the data offloading is conducted only when an MN

opportunistically meets WiFi APs. That is, MNs connect to WiFi networks whenever there exists available WiFi coverage and then migrate their traffic to a WiFi network. The on-the-spot WiFi offloading technique is the simplest WiFi offloading technique, and it is provided in most of the current smartphone platforms.

- *Delayed WiFi offloading*: In the delayed WiFi offloading technique, the data transfer is delayed with the expectation of future AP contacts. That is, an MN delays data traffic of delay-tolerant applications (e.g., file downloads) up to the predefined delay threshold until a WiFi network becomes available. During the delay threshold, if a WiFi network becomes available, the MN immediately starts the pending data session through the WiFi network.

Figure 1 describes the WiFi offloading scenario in heterogeneous networks where WiFi APs are sparsely deployed within cellular coverage. The detailed procedures of each WiFi offloading technique can be described as follows. Consider a scenario where an MN attempts to establish a data session to download/upload a file from/to the Internet at time t . It is assumed that a cellular network is always available, whereas a WiFi network is only available when the MN is close enough to the WiFi AP. If any WiFi network is available at t , in both techniques, the MN establishes the data session only through the WiFi network although a cellular network is available. On the other hand, if a WiFi network is not available at t , in on-the-spot WiFi offloading, the MN uses a cellular network immediately. While for delayed WiFi offloading, the MN defers the data session up to a predefined delay bound D . During the delay bound, if a WiFi network becomes available, owing to the MN's mobility, the MN immediately starts the pending data session through a WiFi network. Therefore, the additional traffic can be offloaded to the WiFi network.

Table 1 (Lee et al. 2013) shows WiFi offloading efficiency for the on-the-spot and delayed WiFi offloading techniques, which are obtained based on WiFi connectivity 97 users for



WiFi Offloading in WiFi-Cellular Networks, Fig. 1 WiFi offloading scenario in heterogeneous networks

WiFi Offloading in WiFi-Cellular Networks, Table 1 WiFi offloading efficiency (Lee et al. 2013)

Offloading scheme	On-the-spot	100 s delay	1 h delay
Offloading efficiency/gains	65%	+(2–3%)	+29%

2 weeks. For the on-the-spot WiFi offloading, WiFi can offload about 65% of the total cellular data traffic without any delay. Meanwhile, it can be seen that the delayed WiFi offloading can offload more traffic at the cost of delay. For 100 s and 1 h delay, the achievable gains of the delayed WiFi offloading are 2–3% and 29%, respectively.

From the fields of academia, there have been several studies on WiFi offloading. The feasibility of WiFi offloading was evaluated in Mota et al. (2013) and Dias et al. (2015). Mota et al. (2013) derived the 3G and WiFi coverage through several bus routes in Paris to investigate how users and network operators can benefit from the existing WiFi infrastructure. On the other hand, Dias et al. (2015) investigated the feasibility of video streaming offloading from cellular networks to WiFi ad hoc networks.

To motivate users to participate in the delayed WiFi offloading, several incentive mechanisms were proposed in the literature (Cai et al. 2015; Zhuo et al. 2014; Lee et al. 2014; Gao et al. 2014). Cai et al. (2015) modeled an incentive mechanism by using a two-stage Stackelberg game. In the first stage, a network operator announces a uniform reward to users. After that, in the second stage, users decide whether to delay the offloading or not. Zhuo et al. (2014) proposed an incentive framework based on the reverse auction to capture dynamic characteristics on users' delay tolerance. Moreover, the authors illustrated how to predict the offloading potential of users by means of stochastic analysis. Lee et al. (2014) investigated economic gains of the delayed WiFi offloading from the perspectives of users and network operators. To this end, the authors formulated a two-stage sequential game where network operators control the price and users are price-takers. Gao et al. (2014) introduced a one-to-many bargaining game among network operators and WiFi access point owners and analyzed the interaction for the amount of data offloading and the respective compensations.

Meanwhile, some researchers investigated how to offload data from cellular networks to

WiFi networks in vehicular ad hoc networks (VANETs) (Wang and Wu 2016; Cheng et al. 2016). Wang and Wu (2016) unified the downloading cost and delay in terms of user's satisfaction and then proposed an adaptive algorithm to maximize the satisfaction. In this algorithm, a vehicle estimates the data size depending on the application type and makes a decision whether to download data through cellular networks or not. Cheng et al. (2016) suggested two WiFi offloading mechanisms (i.e., auction game-based offloading (AGO) and congestion game-based offloading (CGO)). Also, a prediction method for the WiFi offloading potential and access cost within a certain future time was proposed based on the semi-Markov process. Several offloading mechanisms through cooperative downloading for VANETs were introduced in Wu et al. (2015) and Deng et al. (2016). Wu et al. (2015) introduced an incentive-driven cluster-based content distribution mechanism to encourage cooperation. Similarly, Deng et al. (2016) proposed a prior response incentive mechanism (PRIM) to stimulate vehicles to participate in cooperative downloading.

The performance of the WiFi offloading was analyzed in Suh et al. (2016) and Mehmeti and Spyropoulos (2013). Suh et al. (2016) developed analytical models on the WiFi offloading efficiency, which is defined as the ratio of the amount of offloaded data to the total amount of data, and elaborated on the performance of WiFi offloading under diverse environments. Mehmeti and Spyropoulos (2013) proposed a queueing model as a function of WiFi availability, user mobility, and traffic load. However, in these works, how to optimize the delay timer for performance improvement was not investigated.

Few works investigated how to set the delay timer (Mehmeti and Spyropoulos 2014; Ko et al. 2017). Mehmeti and Spyropoulos (2014) proposed a queueing model for the delayed WiFi offloading, where an optimization problem for minimizing download time with cost constraints is formulated to choose the delay timer. Ko et al.

(2017) investigated the effect of the dual connectivity technology on the delayed offloading, and based on the results, they derived the optimal delay timer that can minimize the monetary cost while limiting the outage probability.

For standardization efforts, 3GPP has specified the access network discovery and selection function (ANDSF) that resides within evolved packet core (EPC) and controls offloading between 3GPP and non-3GPP access networks (e.g., WiFi) (3GPP TS 24.312 v. 11.5.0 2012a; 3GPP TS 24.312 v. 11.5.0 2012b). The key functionalities of ANDSF are to assist UEs to discover the available non-3GPP access networks in their proximity and to manage the connections to the non-3GPP access networks by prioritizing them based on operator policies.

Key Application

WiFi offloading techniques are considered as the most promising approaches that mobile operators can take to offload their mobile data traffic into WiFi networks.

Cross-References

- [Delay-Tolerant Network Routing](#)
- [Edge Computing](#)

References

- 3GPP TS 24.312 v. 11.5.0 (2012a) Access network discovery and selection function (ANDSF) management object (MO), Dec 2012
- 3GPP TS 24.312 v. 11.5.0 (2012b) Architecture enhancements for non-3GPP accesses, Dec 2012
- Cai S, Duan L, Wang J, Zhou S, Zhang R (2015) Incentive mechanism design for delayed WiFi offloading. In: Proceedings of the international conference on communications (ICC) 2015, June 2015
- Cheng N, Lu N, Zhang N, Zhang X, Shen X, Mark JW (2016) Opportunistic Wifi offloading in vehicular environment: a game-theory approach. *IEEE Trans Intell Transp Syst* 17(7):1944–1955

- Cisco (2017) Visual networking index: global mobile data traffic forecast update, 2016–2021. White Paper, Mar 2017
- Deng G, Li F, Wang L (2016) Cooperative downloading in VANETs-LTE heterogeneous network based on named data. In: Proceedings of the workshop on name-oriented mobility: architecture, algorithms and applications, Apr 2016
- Dias E, Vargas E, Farias M, Carvalho M (2015) Feasibility of video streaming offloading via connection sharing from LTE to WiFi Ad Hoc networks. In: Proceedings of the international workshop on telecommunications (IWT) 2015, June 2015
- Gao L, Iosifidis G, Huang J, Tassiulas L, Li D (2014) Bargaining-based mobile data offloading. *J Sel Areas Commun* 32(6):1114–1125
- Jung B, Song N, Sung D (2014) A network-assisted user-centric WiFi-offloading model for maximizing per-user throughput in a heterogeneous network. *IEEE Trans Veh Technol* 63(4):1940–1945
- Juniper Research (2013) Data offload connecting intelligently. White Paper, May 2013
- Ko H, Lee J, Pack S (2017) Performance optimization of delayed WiFi offloading in heterogeneous networks. *IEEE Trans Veh Technol* 66(10):9436–9447
- Lee K, Lee J, Yi Y, Rhee I, Chong S (2013) Mobile data offloading: How much can WiFi deliver? *IEEE/ACM Trans Netw* 21(2):536551
- Lee J, Yi Y, Chong S, Jin Y (2014) Economics of WiFi offloading: trading delay for cellular capacity. *IEEE Trans Wirel Commun* 13(3):1540–1554
- Mehmeti F, Spyropoulos T (2013) Performance analysis of “on-the-spot” mobile data offloading. In: Proceedings of the IEEE global communications conference (GLOBECOM) 2013, Dec 2013
- Mehmeti F, Spyropoulos T (2014) Is it worth to be patient? Analysis and optimization of delayed mobile data offloading. In: Proceedings of the IEEE international conference on computer communications (INFOCOM) 2014, Apr/May 2014
- Mota V, Macedo D, Ghamri-Doudane Y, Nogueira J (2013) On the feasibility of WiFi offloading in urban areas: the Paris case study. In: Proceedings of the IFIP wireless days (WD) 2013, Nov 2013
- Suh D, Ko H, Pack S (2016) Efficiency analysis of WiFi offloading techniques. *IEEE Trans Veh Technol* 65(5):3462–3473
- Wang N, Wu J (2016) Opportunistic Wifi offloading in a vehicular environment: waiting or downloading now? In: Proceedings of the IEEE international conference on computer communications (INFOCOM) 2016, Apr 2016
- Wu C, Gerla M, Mastrorade N (2015) Incentive Driven LTE Content Distribution in VANETs. In: Proceedings of the annual mediterranean ad hoc networking workshop (MED-HOC-NET) 2015, June 2015
- Zhuo X, Gao W, Cao G, Hua S (2014) An incentive framework for cellular traffic offloading. *IEEE Trans Mobile Comput* 13(3):541–555

Wi-Fi Protected Access (WPA)

Sheng Zhong

Nanjing University, Nanjing, China

Synonyms

[The draft IEEE 802.11i standard](#)

Definition

Wi-Fi Protected Access (WPA) is a security standard designed for wireless networks. WPA implements most of the IEEE 802.11i standard and was an alternative to previous security standard before 802.11i was developed, so it is sometimes referred to as the *draft IEEE 802.11i* standard.

Historical Background

WPA was developed by the Wi-Fi Alliance to provide better security than Wired Equivalent Privacy (WEP), the previous Wi-Fi security standard. WPA became available in 2003, and it was intended as a transitional measure for the full IEEE 802.11i standard (also referred to as WAP2), which became available in 2004. WAP2 may not work with some older network cards, while WPA can work well with these older network cards. Therefore, WPA and WPA2 are concurrent security standards.

There are typically two versions of WPA based on the differences in target users:

- WPA-Personal mode, also referred to as WPA-PSK, is designed for home and small office networks, which makes every user under the same wireless router use the same key called pre-shared key (PSK).
- WPA-Enterprise mode, also referred to as WPA-EAP, uses more stringent 802.1x authentication with the Extensible Authen-

tication Protocol (EAP). It needs to use an authentication server to distribute different keys to various end users.

Except for authentication, WPA also offers tremendous improvements in encryption and data integrity compared to WEP.

- **Encryption.** WEP only has a 64-bit or 128-bit encryption key which is fixed and has to be manually entered on wireless access points (APs). WPA adopts a so-called Temporary Key Integrity Protocol (TKIP), which dynamically produces a new 128-bit key for each data packet. It defeats well-known related-key attack against WEP.
- **Data integrity.** WPA also adopts a message integrity check, which is designed to prevent an attacker from altering and resending data packets, and it replaces the cyclic redundancy check (CRC) used by WEP. CRC's main drawback is that it does not provide a sufficiently strong data integrity guarantee for the packets it handles. Well-tested message authentication codes can solve these problems, but they require too much computation to be used on old network cards. WPA uses message integrity check algorithm TKIP to verify the integrity of the packets, which needs less computation but offers stronger performance than CRC.

TKIP was designed to be used with older WEP devices. However, researchers did discover a flaw in TKIP that can be used to achieve a certain attack. This caused the replacement of TKIP by CCMP (sometimes called AES-CCMP) encryption protocol in WPA2, which provides additional security.

The latest security standard for wireless networks is WPA3, which was announced as a replacement to WPA2 by the Wi-Fi Alliance in January 2018. WPA3 uses 192-bit key and individualized encryption for each user. The Wi-Fi Alliance also claims that WPA3 will mitigate security issues posed by weak passwords and simplify the process of setting up devices with no display interface.

Security Issues

There are a series of security issues in WPA.

1. **Weak passwords.** If users rely on weak passwords, WPA-PSK is still vulnerable to password cracking attacks. Brute forcing of simple passwords can be attempted using the Aircrack Suite (Tsitroulis et al. 2014). WPA3 solves this problem by requiring interaction with the infrastructure for each guessed password, so that the infrastructure may place temporal limits on the number of guesses.
2. **A lack of forward secrecy.** WPA-PSK doesn't provide forward secrecy, meaning that once an adversary discovers the pre-shared key, they can potentially decrypt all packets encrypted using that PSK. This also means an attacker can silently capture and decrypt others' packets if an access point is provided free of charge at a public place, because its password is usually shared to anyone in that place. Therefore, it's safer to use Transport Layer Security (TLS) or something similar on top of that for transferring any sensitive data.
3. **WPA packet spoofing and decryption.** The Beck-Tews attack (Tews and Beck 2009) could decrypt short packets with mostly known content, such as ARP messages, and allowed injection of 3 to 7 packets of at most 28 bytes. Halvorsen and others (Halvorsen et al. 2009) show how to modify the Beck-Tews attack to allow injection of 3 to 7 packets having a size of at most 596 bytes. The Vanhoef-Piessens attack (Vanhoef and Piessens 2013) significantly improved the Beck-Tews attack, which can inject an arbitrary amount of packets and decrypt arbitrary packets sent to a client. The vulnerability of TKIP is fixed by AES-based CCMP adopted by WPA2.
4. **KRACK attack.** The KRACK attack published in 2017 is believed to affect all variants of WPA and WPA2 (Vanhoef and Piessens 2017). Software patches can resolve the vulnerability but are not available for all devices.

Key Applications

WPA ensures that non-authorized users cannot access the wireless networks and ensure users who legitimately gain access to the networks cannot hack other devices on the same hotspot. WPA is still a configuration option upon a wide variety of wireless routing devices. A survey in 2013 showed that 71% still allowed usage of WPA and 19% exclusively supported WPA (Vanhoeft and Piessens 2013).

Cross-References

► [WiFi Offloading in WiFi-Cellular Networks](#)

References

- Halvorsen FM, Haugen O, Eian M, Mjøltnes SF (2009) An improved attack on TKIP. In: Nordic conference on secure IT systems. Springer, pp 120–132
- Tews E, Beck M (2009) Practical attacks against WEP and WPA. In: Proceedings of the second ACM conference on wireless network security. ACM, pp 79–86
- Tsitroulis A, Lampoudis D, Tsekles E (2014) Exposing WPA2 security protocol vulnerabilities. *Int J Inf Comput Secur* 6(1):93–107
- Vanhoeft M, Piessens F (2013) Practical verification of WPA-TKIP vulnerabilities. In: Proceedings of the 8th ACM SIGSAC symposium on information, computer and communications security. ACM, pp 427–436
- Vanhoeft M, Piessens F (2017) Key reinstallation attacks: forcing nonce reuse in WPA2. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. ACM, pp 1313–1328

Willingness-to-Pay

► [Preferences and Utility Functions](#)

Wireless 4G/LTE Key Technologies

► [Key Technologies in 4G/LTE Network](#)

Wireless Access in Vehicular Environment

Ruobing Jiang¹ and Yanmin Zhu²

¹Ocean University of China, Qingdao, China

²Shanghai Jiao Tong University, Shanghai, China

Synonyms

[Dedicated short-range communication \(DSRC\)](#); [ETSI ITS-G5 standard](#); [IEEE 1609 standards set](#); [IEEE Std 802.11p](#); [Vehicle-to-Everything \(V2X\) Communication](#); [Vehicular communication system](#)

Definition

A wireless access in vehicular environments (WAVE) system provides interoperable, efficient, and reliable radio communications in support of applications offering safety and convenience in an intelligent transportation system (ITS). In many of the ITS application scenarios, the WAVE system is designed to provide vehicles with the direct connectivity to other vehicles (V2V), to roadside (V2R), or to infrastructures (V2I) through dedicated short-range communications (DSRC). In the USA, DSRC is used to refer to radio spectrum in the 5.9 GHz band or technologies associated with WAVE. While outside the USA, DSRC may refer to a distinct radio technology operating at 5.8 GHz.

The architecture and operations of a WAVE system are based on the IEEE 1609 series of standards (IEEE 2014a) and the IEEE 802.11p standard, which is now incorporated in IEEE Std 802.11-2012. IEEE Std 1609.0 is a comprehensive guide describing the WAVE architecture and services necessary for a WAVE system. This guide is meant to be used in conjunction with IEEE Std 1609.4 (multichannel operations), IEEE Std 1609.3 (networking services), IEEE Std 1609.2 (Security Services for Applications and Management Messages),

IEEE Std 1609.11 (Over-the-Air Electronic Payment Data Exchange Protocol for ITS), IEEE Std 1609.12 (Identifier Allocations), IEEE Std 802.11 (operations outside the context of a basic service set), and other WAVE standards that may be developed to specify higher layer, application, or features.

A WAVE device is a device compliant to the IEEE 1609 standards and IEEE Std 802.11 and supporting the information exchange with other WAVE devices. Typical WAVE devices are onboard units (OBUs) mounted in onboard equipment in vehicles and roadside units (RSUs) that generally operate when stationary, which can be installed in roadside poles, traffic lights, road signs, or roadside electronic cabinets. WAVE devices adopt orthogonal frequency-division multiplexing (OFDM) and achieve a communication range of approximate 1,000 ft.

Historical Background

Intelligent transportation systems (ITSs) have been developed throughout the world since the early 1990s to improve transportation safety, traffic efficiency, and user comfort and convenience. The coordination among moving vehicles and roadside infrastructures enables high-performance traffic management, pollution reduction, and energy conservation as well.

To implement such intelligent transportation systems, wireless communications and networking have been recognized, from the beginning, as a cornerstone and play an essential role in the evolution of such systems. The IEEE has developed the WAVE architecture and associated standards to promote and standardize the widely application of ITSs. A brief context for the IEEE WAVE standards and related activities are introduced as follows.

Back in 1991, the US Congress mandated the creation of the US ITS program called Intelligent Vehicle Highway Systems (IVHS) in the 1991 Intermodal Surface Transportation Efficiency Act (ISTEA). The IVHS program is administered by the US Department of Transportation (DOT) and supported by the

nonprofit organization, Intelligent Transportation Society of America (ITSA). The program aimed at applying advanced concepts and technologies in areas of communications, navigation, and information systems to reduce traffic congestion, increase transportation efficiency, enhance mobility, improve highway safety, and reduce harm to the environment caused by automobiles (Sweeney 1993). The USDOT renamed the IVHS program into ITS program to clarify the multimodal intent.

In the mid-1990s, a national systems architecture, i.e., the National ITS Architecture, was developed by the ITS program office. Furthermore, the related ITS standards were established by USDOT in 1996 to encourage the widespread use of ITS technologies in the USA.

In 1997, the Transportation Equity Act for the twenty-first century (TEA-21) was passed, which retained ISTEA's essential features while boosting investments to promote the deployment of the USDOT's ITS program.

In 1999, the US Federal Communications Commission (FCC) allocated 75 MHz in the 5.850–5.925 GHz band to DSRC uses (FCC 2003), in response to the petition of the ITS America in 1997.

In 2004, the FCC published the DSRC Report and Order, FCC 03-324 (FCC 2004), which established standard licensing and service rules in the ITS Radio Service in the 5.9 GHz band. The published FCC 03-324 adopted the standard for the physical and MAC layers (ASTM's E2213-03 ASTM 2003), which was developed by the American Society for Testing and Materials (ASTM) based on IEEE 802.11.

In November 2004, the IEEE 802.11 Task Group p was formed to develop an amendment to the IEEE 802.11 standard to include WAVE. The specifications for higher layers in the protocol suite were developed by another IEEE team, working group 1609.

In 2006 and 2007, a set of IEEE 1609 standards, as well as IEEE Std 802.11p, were managed for trial use to demonstrate the performance of WAVE architecture and protocols. The trial-used IEEE 1609 standards include IEEE Std 1609.4–2006 (IEEE 2006b),

IEEE Std 1609.3–2007 (IEEE 2007), IEEE Std 1609.2–2006 (IEEE 2006a), and IEEE Std 1609.1–2006 (IEEE 2006c).

Since 2010, a set of full-use IEEE 1609 standards have been published. The currently active full-use standards include IEEE 1609.0-2013 (IEEE 2014a), IEEE 1609.2-2016 (IEEE 2013, 2016a), IEEE 1609.3-2016 (IEEE 2010, 2012b, 2014b, 2016b), IEEE 1609.4-2016 (IEEE 2011b, 2016d), IEEE 1609.11-2010 (IEEE 2011a), and IEEE 1609.12-2016 (IEEE 2012a, 2016c).

Foundations

Spectrum Allocation

The 75 MHz bandwidth allocated by the US FCC for DSRC usage is divided into seven 10 MHz operation channels, and the remaining 5 MHz bandwidth is reserved.

As illustrated in Fig. 1, the spectrum from 5.855 to 5.865 GHz is defined as Channel 172, while the spectrum from 5.865 to 5.875 GHz is defined as Channel 174, and so on, until Channel 184. Among all the defined seven channels, Channel 178 is used as Control Channel (CCH), while the 20-MHz Channel 175, covering Channel 174 and Channel 176, and Channel 181, merged by Channel 180 and 182, are specified as service channel (SCH). The rest Channel 172 and Channel 184 are kept for public safety applications. Specifically, according to FCC 06-110 document (FCC 2006), Channel 172 is designated for “vehicle-to-vehicle safety communications for accident avoidance and mitigation and safety of life and property applications.” Channel 184 is specified for “high-power, longer-distance

communications to be used for public safety applications involving safety of life and property, including road intersection collision mitigation.”

The IEEE WAVE standards specify two kinds of radio channel, i.e., a control channel and several service channels. The CCH 178 is used only for system management messages and WAVE Short Message Protocol (WSMP) messages. The WSMP is a protocol for rapid exchange of the messages subject to intermittent radio connectivity, which will be described in the next subsection. To reducing the delivery delay, WAVE Short Messages (WSMs) are transmitted with controllable physical characteristics and thus could be transmitted with minimal channel capacity. As a result, WSMs are allowed to use both CCH and SCHs.

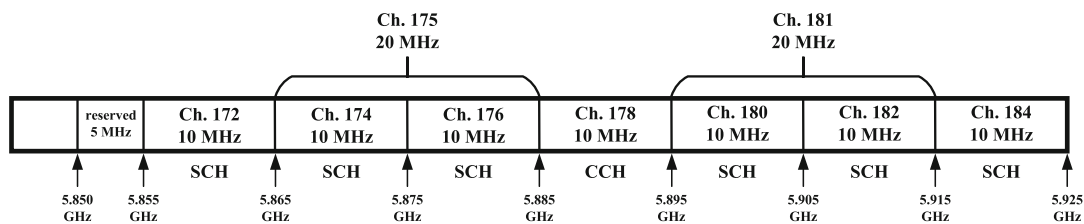
The service channels are used for application data transfers as well as IPv6 traffic.

Protocol Stack

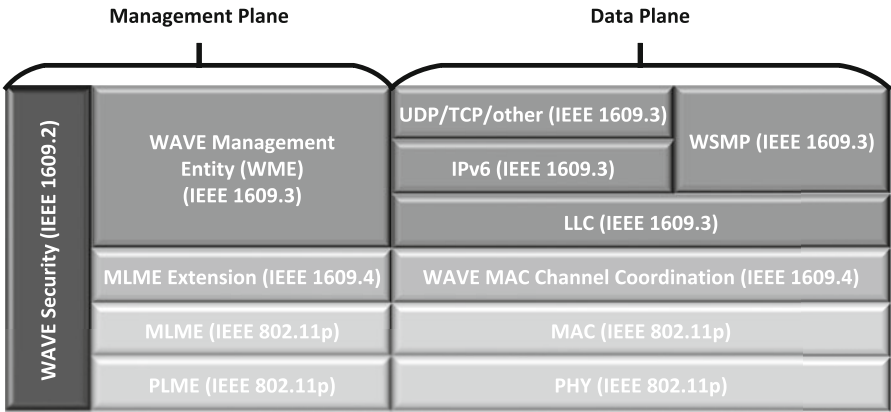
The WAVE protocol stack and the mapping of these protocol elements to IEEE WAVE standards are illustrated in Fig. 2. A data plane with protocols carrying higher layer information and a management plane for security and management functions are defined.

The full-use IEEE 1609 standards and IEEE Std 802.11p, which has been incorporated in IEEE 802.11-2012, collectively define the WAVE protocol stack.

IEEE Std 1609.3 specifies two data plane protocol stacks, the Internet Protocol Version 6 (IPv6) standard and the WSMP, which are both based on the common set of lower layer protocols including physical (PHY), medium access control (MAC) and logical link control (LLC). IEEE Std 1609.3 specifies the use of two Ethertype



Wireless Access in Vehicular Environment, Fig. 1 WAVE channels allocated by the U.S. FCC



Wireless Access in Vehicular Environment, Fig. 2 WAVE protocol stack

values, IPv6 or WSMP, in the Ethertype field in the LLC header. IEEE Std 1609.3 also specifies a WAVE Management Entity (WME).

IEEE Std 802.11p, as an amendment to IEEE 802.11, specifies several physical layers (PHYs) and one medium access control (MAC) sublayer.

IEEE Std 1609.4 specifies multichannel operations in the data plane and extensions to the IEEE 802.11 MAC sublayer management entity (MLME).

Security services for WAVE applications and management messages are defined by IEEE Std 1609.2.

IEEE Std 1609.11 specifies the first application-level protocol for Over-the-Air Electronic Payment Data exchange in ITS. IEEE Std 1609.12 standardized the used identifiers and the allocated values of the identifiers in WAVE systems.

IEEE Std 1609.0 is a guide indicating the necessary services and architecture for WAVE system.

Key Applications

The applications of WAVE systems can be classified according to their primary purposes, such as safety, efficiency, comfort, on-the-road services, or urban sensing.

- **Safety applications:** One of the most important categories of WAVE applications

is to enhance transportation safety through direct V2V or V2I communications. Example applications include forward collision warning, intersection violation warning, icy road warning, and blind spot warning. IEEE Std 1609.2 WAVE Security Services are employed to sign and authenticate such safety messages to avoid encryption when broadcasting, and the sign messages are sent using WSMP to promote bandwidth efficiency.

- **Efficiency applications:** This category of applications aim to improve the traffic efficiency of the whole urban transportation system. Representative examples are traffic lights scheduling, route planning and navigation, and electronic fee collection (EFC). The general objective of the EFC application is to conduct tolling transaction reliably and securely while vehicles pass at highway speeds. The tolling transaction should be accomplished with unambiguous association of the payer’s physical vehicle for compatibility with revenue protection violation detection and enforcement systems. The EFC transaction technology is specified by IEEE Std 1609.3, IEEE Std 1609.4 and IEEE Std 1609.11.
- **On-the-road services:** The service information received by drivers on-the-road makes the trips more comfortable and enjoyable. Such information includes weather information,

point-of-interest locations, advertisement of on-going sales, and available parking lot.

- **Urban sensing:** Based on the V2V and V2I communications, the WAVE systems also support urban monitoring and sensing when vehicles are equipped with onboard sensors. The photos and videos of road conditions, driving speeds of vehicles, or collision or emergency alerts could be shared among vehicles or through Internet for large-scale urban monitoring and efficient sensing data collection.

Cross-References

- [Capacity of Wireless Ad Hoc Networks](#)
- [Evolution of Vehicular Networks in the Transition from 3G to 4G Systems](#)
- [Quality of Service in IEEE 802.11 Networks](#)

References

- ASTM (2003) Standard specification for telecommunications and information exchange between roadside and vehicle systems—5 GHz band dedicated short range communications (DSRC) medium access control (MAC) and physical layer (PHY) specifications. Technical report, ASTM E2213-03
- FCC (2003) Amendment of parts 2 and 90 of the commission's rules to allocate the 5.850–5.925 GHz band to mobile service for dedicated short-range communications of intelligent transportation services. Technical report, FCC ET Docket No. 98-95, RM-9096
- FCC (2004) Amendment of the commission's rules regarding dedicated short-range communication services in the 5.850–5.925 GHz band (5.9 GHz band). Technical report, FCC 03–324
- FCC (2006) Amendment of the commission's rules regarding dedicated short-range communication services in the 5.850–5.925 GHz band (5.9 GHz band). Technical report, FCC 06–110
- IEEE (2006a) IEEE trial-use standard for wireless access in vehicular environments – security services for applications and management messages. IEEE Std 16092-2006, pp 1–105. <https://doi.org/10.1109/IEEESTD.2006.243731>
- IEEE (2006b) IEEE trial-use standard for wireless access in vehicular environments (wave) – multi-channel operation. IEEE Std 16094-2006, pp 1–82. <https://doi.org/10.1109/IEEESTD.2006.254109>
- IEEE (2006c) Trial-use standard for wireless access in vehicular environments (wave) – resource manager. IEEE Std 16091-2006, pp 1–71. <https://doi.org/10.1109/IEEESTD.2006.246485>
- IEEE (2007) IEEE trial-use standard for wireless access in vehicular environments (WAVE) – networking services. IEEE Std 16093–2007, pp 1–99. <https://doi.org/10.1109/IEEESTD.2007.353212>
- IEEE (2010) IEEE standard for wireless access in vehicular environments (wave) – networking services. IEEE Std 16093-2010 (Revision of IEEE Std 16093-2007), pp 1–144. <https://doi.org/10.1109/IEEESTD.2010.5680697>
- IEEE (2011a) IEEE standard for wireless access in vehicular environments (wave)– over-the-air electronic payment data exchange protocol for intelligent transportation systems (its). IEEE Std 160911-2010, pp 1–62. <https://doi.org/10.1109/IEEESTD.2011.5692959>
- IEEE (2011b) IEEE standard for wireless access in vehicular environments (WAVE)-multi-channel operation. IEEE Std 16094-2010 (Revision of IEEE Std 16094-2006), pp 1–89. <https://doi.org/10.1109/IEEESTD.2011.5712769>
- IEEE (2012a) IEEE standard for wireless access in vehicular environments (wave) – identifier allocations. IEEE Std 160912-2012, pp 1–20. <https://doi.org/10.1109/IEEESTD.2012.6308691>
- IEEE (2012b) IEEE standard for wireless access in vehicular environments (wave)–networking services corrigendum 1: miscellaneous corrections. IEEE Std 16093-2010/Cor 1-2012 (Corrigendum to IEEE Std 16093-2010), pp 1–19
- IEEE (2013) IEEE standard for wireless access in vehicular environments security services for applications and management messages. IEEE Std 16092-2013 (Revision of IEEE Std 16092-2006), pp 1–289. <https://doi.org/10.1109/IEEESTD.2013.6509896>
- IEEE (2014a) IEEE guide for wireless access in vehicular environments (WAVE) - architecture. IEEE Std 16090-2013, pp 1–78. <https://doi.org/10.1109/IEEESTD.2014.6755433>
- IEEE (2014b) IEEE standard for wireless access in vehicular environments (wave) – network systems corrigendum 2: miscellaneous corrections. IEEE Std 16093-2010/Cor 2-2014 (Corrigendum to IEEE Std 16093-2010), pp 1–33. <https://doi.org/10.1109/IEEESTD.2014.6998915>
- IEEE (2016a) IEEE standard for wireless access in vehicular environments–security services for applications and management messages. IEEE Std 16092-2016 (Revision of IEEE Std 16092-2013), pp 1–240. <https://doi.org/10.1109/IEEESTD.2016.7426684>
- IEEE (2016b) IEEE standard for wireless access in vehicular environments (WAVE) – networking services. IEEE Std 16093-2016 (Revision of IEEE Std 16093-2010), pp 1–160. <https://doi.org/10.1109/IEEESTD.2016.7458115>
- IEEE (2016c) IEEE standard for wireless access in vehicular environments (wave) – identifier allocations. IEEE Std 160912-2016 (Revision of IEEE

- Std 160912-2012), pp 1–21. <https://doi.org/10.1109/IEEESTD.2016.7428792>
- IEEE (2016d) IEEE standard for wireless access in vehicular environments (WAVE) – multi-channel operation. IEEE Std 16094-2016 (Revision of IEEE Std 16094-2010), pp 1–94. <https://doi.org/10.1109/IEEESTD.2016.7435228>
- Sweeney LE (1993) An overview of intelligent vehicle highway systems (IVHS). In: Proceedings of WESCON '93, pp 258–262. <https://doi.org/10.1109/WESCON.1993.488444>

Wireless Ad Hoc Networks

- [Live Video Streaming in Wireless P2P Networks](#)

Wireless Big Data

Peng Li
School of Computer Science and Engineering,
The University of Aizu, Aizu-Wakamatsu,
Fukushima, Japan

Synonyms

[Wireless big data](#)

Definitions

Wireless big data describe massive data (e.g., wireless signaling data, device logs, browser history, and payment records) generated by wireless networks (e.g., cellular networks, wireless sensor networks, and wireless ad hoc networks), as well as related technologies for data collection and analytics.

Historical Background

Similar to the traditional concept of big data, wireless big data can be also described by four “V”: volume, variety, velocity, and value.

In addition, wireless big data show unique characteristics of large-scale distribution and strong heterogeneity, which impose additional challenges in data collection and analytics.

Wireless big data are generated everywhere in wireless networks, and their collection incurs high communication cost. In the early period, a natural approach for reducing communication cost is to sample a subset of data. In particular, compressive sampling theory (Donoho 2006) has been successfully applied in wireless sensor networks (Luo et al. 2009). It can extract global information by collecting data from a subset of sensors. Later, thanks to the significant improvement of network capacity, it becomes affordable to collect all data to avoid information missing. However, heavy network traffic generated in wireless big data collection still exists, and it can affect other applications running over wireless networks. For example, some network service providers collect system logs generated by cellular network infrastructures (e.g., base stations and switches) for network diagnosis. These traffics occupy a portion of network bandwidth, and they should be well scheduled to minimize the influence to mobile user traffic.

Meanwhile, wireless big data show heterogeneity because they are generated by various devices and applications. They have different formats and are collected with different frequencies. To provide a unified data representation, a tensor model has been proposed to represent the unstructured, semistructured, and structured data generated in wireless networks (Kuang et al. 2014). He et al. (2016) have introduced a unified data model based on the random matrix theory and machine learning.

Wireless big data can be analyzed by techniques (e.g., data mining and pattern recognition) that are also adopted by traditional big data analytics. In addition, some wireless big data can be expressed in spatial-temporal dimensions that impose additional challenges. Xu et al. (2016) design a time series analysis approach that can decompose large-scale mobile traffic into regularity and randomness components. Then, the traffic patterns can be predicted based on regularity components. In recent years, there is a growing

trend of applying machine learning techniques, especially deep learning, in wireless big data analytics. Alsheikh et al. (2016) present an overview and brief tutorial on deep learning in mobile big data analytics, as well as a scalable learning framework over Apache Spark. Yang (2016) investigates various learning methodologies for wireless big data analytics. Qian et al. (2017) survey recent research results of wireless big data analytics.

Key Applications

Wireless big data can be applied in many scenarios, such as user mobility modeling, traffic analysis, network planning, privacy, and security enhancement. Some key applications are listed as follows.

- **User Mobility Analysis:** Wireless big data generated by mobile phones have great potential for human mobility analysis. By mining wireless big data, the human mobility can be understood, predicted, or even controlled. For instance, by recording and analyzing the time spent at different locations for different individuals, it can be robustly identified whether an individual is at home or at office, which can improve the efficiency urban cellular network planning. Moreover, individual mobility and mobile traffic are closely linked to the social ecology, which helps achieve a better understanding of human behavior.
- **Social Networks:** Wireless big data also exist in various mobile social networks. Massive high dimensional personal data are collected from mobile users by the cloud. Analyzing these data can predict individual activities and provide personalized recommendation. A typical example is video recommendation. The user context data can be analyzed by a recommendation model, so that appropriate videos can be recommended for users. Meanwhile, recommendation policy can be improved according to user feedbacks.
- **Security and Privacy:** Although wireless big data can help understand individual behavior more easily, the privacy and security issues are inevitable. Wireless big data containing various privacy information, such as user locations, interests, and habits, which would be easily obtained by malicious users due to the broadcast nature of wireless transmission. Therefore, many applications concentrate on protecting privacy and security of wireless big data. For instance, using the method of distorting time, the interest and mobility habits can be concealed for achieve anonymous mobile data set.

Cross-References

► [Big Data in 5G](#)

References

- Alsheikh MA, Niyato D, Lin S, Tan HP, Han Z (2016) Mobile big data analytics using deep learning and apache spark. *IEEE Netw* 30(3):22–29. <https://doi.org/10.1109/MNET.2016.7474340>
- Aron M, Brown B (2001) The politics of nature. In: Smith J (ed) *The rise of modern genomics*, 3rd edn. Wiley, New York, pp 234–295
- Donoho DL (2006) Compressed sensing. *IEEE Trans Inf Theory* 52(4):1289–1306. <https://doi.org/10.1109/TIT.2006.871582>
- He Y, Yu FR, Zhao N, Yin H, Yao H, Qiu RC (2016) Big data analytics in mobile cellular networks. *IEEE Access* 4:1985–1996. <https://doi.org/10.1109/ACCESS.2016.2540520>
- Kuang L, Hao F, Yang LT, Lin M, Luo C, Min G (2014) A tensor-based approach for big data representation and dimensionality reduction. *IEEE Trans Emerg Top Comput* 2(3):280–291. <https://doi.org/10.1109/TETC.2014.2330516>
- Luo C, Wu F, Sun J, Chen CW (2009) Compressive data gathering for large-scale wireless sensor networks. In: *Proceedings of the 15th annual international conference on mobile computing and networking*. ACM, New York, *MobiCom' 09*, pp 145–156. <https://doi.org/10.1145/1614320.1614337>
- Qian L, Zhu J, Zhang S (2017) Survey of wireless big data. *J Commun Inf Netw* 2(1):1–18. <https://doi.org/10.1007/s41650-017-0001-2>
- Xu F, Lin Y, Huang J, Wu D, Shi H, Song J, Li Y (2016) Big data driven mobile traffic understanding and forecasting: a time series approach. *IEEE Trans*

Serv Comput 9(5):796–805. <https://doi.org/10.1109/TSC.2016.2599878>

Yang C (2016) Learning methodologies for wireless big data networks. Neurocomput 174(PA):431–438. <https://doi.org/10.1016/j.neucom.2015.04.111>

Wireless Blockchain Networks

- [Mining Task Offloading in Wireless Blockchain Networks](#)

Wireless Body Area Networks

- [60 GHz in WBAN and mm-Wave Technologies](#)

Wireless Caching

- [Wireless Edge Caching](#)

Wireless Charger Deployment with Communication Constraint

Haipeng Dai, Nan Yu, Alex X. Liu,
Bingchuan Tian, and Guihai Chen
Nanjing University, Nanjing, Jiangsu, China

Synonyms

[Wireless charger placement with communication constraint](#)

Definitions

Communication constraint is that if the distance of placed two chargers is less than or equal to communication radius R_c , then two chargers can communicate with each other; otherwise,

two chargers cannot communicate with each other.

Historical Background

Power transfer (WPT) technology enables the transmission of energy from a power source to an electrical load without any interconnecting cord (Lu et al. 2016). Due to its benefits such as no-wiring, no-contact, and reliability, WPT technology has been utilized in many areas such as public transportation (Mahmoudi et al. 2014), medical apparatuses (RamRakhyani et al. 2011), underwater high-power transfer system (Cheng et al. 2015), and wireless identification and sensing platform. So far, there are more than 300 million WPT commercial products in use.

A wireless charging system typically consists of multiple power transmitters, or called *wireless chargers*, and multiple power receivers, or called *rechargeable devices*. Wireless chargers commonly need to communicate with each other to exchange information like position, time, charging power, and collected data from rechargeable devices for practical purposes such as (1) collaborative scheduling to achieve certain goals such as optimization of overall charging efficiency or electromagnetic radiation safety; (2) data gathering from chargers and/or devices to a certain sink; and (3) configuration updating from a certain sink to chargers and/or devices. These purposes necessitate the need for connectivity of chargers. Besides, comparing with wired communication, wireless communication demonstrates its special advantages for enabling the connectivity of chargers, for instance: (1) it reduces the financial and labor overhead for communication line cost, layout and maintenance; (2) it is more convenient for communication establishment when chargers dynamically join or leave the network; (3) it is more reliable because communication lines expose to cutoff risks in harsh environments such as outdoor and factories; (4) it is much more preferable for mobile chargers (Madhja et al. 2015). For example, unmanned aerial vehicles equipped with wireless chargers can be

used to transmit wireless power to sensor nodes mounted on civil structures such as bridges for monitoring (Mascareñas et al. 2008); chargers can be carried by a boat to charge underwater sensor nodes (Cheng et al. 2015). Motivated by these reasons, Powercast corporation plans to embed wireless communication modules into their wireless charging products.

The problem of Connected *w*ireless Charger *p*lacement (CIRCLE) is introduced. In the considered scenario, there are some directional rechargeable devices with directional antennas distributed on a 2D plane with known positions and orientation angles. To quantify charging efficiency, charging utility for a device is given that is positively correlated with the device's aggregated received power. Suppose there needs to deploy a limited number of directional wireless chargers with directional antennas to some candidate positions on the plane to supply power to the devices. A candidate position can be placed with no more than a fixed number of chargers. The orientation angles of the chargers can be arbitrarily set. The goal is to maximize the overall charging utility for the devices under the constraint that all chargers should be connected.

There exist many works considering the wireless charger placement problem, but none of them take into account the connectivity issue regarding chargers. He et al. (2013) concerned the wireless charger placement problem where static or mobile tags must be guaranteed to receive sufficient power to keep their continuous working. Chiu et al. (2012) exploited the mobility nature of sensor nodes to improve charger deployment. Dai et al. (2016b) considered the electromagnetic radiation safety issues for wireless charger placement. There are also literatures (Dai et al. 2017) considering the directional charging property of chargers and/or devices. Moreover, there are some literatures (Horster et al. 2006; Fusco et al. 2009; Han et al. 2008) considering directional sensor placement. However, most of them are essentially geometric problems, which are fundamentally different from the considered problem and cannot be applied to address CIRCLE. In addition, some other literatures consider the problem of

maximizing a submodular set function with a connectivity constraint (Kuo et al. 2015; Huang et al. 2015; Khuller et al. 2014), which is shown to be mostly related to CIRCLE. Nevertheless, none of them can be directly used to address CIRCLE. The root reason is that CIRCLE adopt the practical directional charging model for both chargers and devices and consider the practical budget constraint for each candidate placement position.

The considered problem has three main technical challenges. The first challenge is that CIRCLE is nonlinear and NP-hard. CIRCLE is nonlinear because finding a connected subgraph in a connected graph toward a certain goal is in essence a combinatorial optimization problem, and the charging utility function is not linear. The second challenge is to design an effective algorithm with a guaranteed approximation ratio to find the placement positions for chargers with connectivity constraint. The third challenge is to select the best orientation angles for chargers from a continuous solution space.

First, a relaxed version of the nonlinear and NP-hard problem CIRCLE is considered by using omnidirectional charging models for chargers and devices, CIRCLE-R, and CIRCLE-R is proved to be a problem of maximizing a submodular set function with connectivity constraint as shown in the following Lemma 1. Thus the first challenge is addressed. Second, a new approximation algorithm to address CIRCLE-R is proposed. Most importantly, CIRCLE-R achieves an approximation ratio that outperforms the state-of-the-art algorithm by at least 1.5 times. Therefore, the second challenge is addressed. Third, to address the original problem CIRCLE, selecting the best orientation angle for each charger can be equivalently considered as a limited number of candidate orientation angles rather than all possible angles in $[0, 2\pi)$ without performance loss, and thus the third challenge is addressed.

Foundations

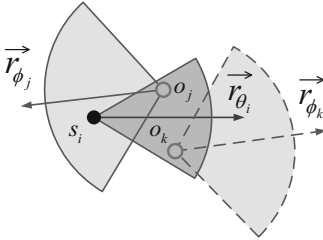
Because of the page limitation, our first algorithm for CIRCLE-R is introduced here, and detailed

analysis and proof are referred to the paper (Yu et al. 2018).

Network Model and Charging Model

Suppose there are N rechargeable devices denoted as $O = \{o_1, \dots, o_N\}$ in a $2D$ area Ω with fixed known positions and orientation angles. There are also a set of homogeneous B directional wireless chargers $S = \{s_1, \dots, s_B\}$ to deploy in the M candidate positions $\mathcal{P} = \{p_1, \dots, p_M\}$ that constitute a connected graph in Ω . A pair of chargers s_i and s_j can communicate with each other if their distance is not larger than R_c . The general directional charging model that is proposed in (Dai et al. 2016a, 2017) (note that the omnidirectional charging model is a special case of the directional charging model by setting the charging angle of chargers to 2π) is adopted here. As Fig. 1 shows, a charger s_i with orientation vector denoted by $\vec{r}_{\theta_i}$ can only supply power for devices in a (power) charging area as a sector region with (power) charging angle A_s and radius D . Similarly, a rechargeable device o_j with orientation vector denoted by $\vec{r}_{\phi_j}$ can only receive nonnegative power in a (power) receiving area as a sector region with (power) receiving angle A_o and radius D . Formally, the charging power from charger s_i to device o_j can be represented as follows.

$$\Pr(s_i, \theta_i, o_j, \phi_j) = \begin{cases} \frac{\alpha}{(\beta + \|\vec{s}_i \vec{o}_j\|)^2}, & 0 \leq \|\vec{s}_i \vec{o}_j\| \leq D, \\ \frac{\vec{s}_i \vec{o}_j \cdot \vec{r}_{\theta_i} - \|\vec{s}_i \vec{o}_j\| \cos(\frac{A_s}{2})}{\|\vec{s}_i \vec{o}_j\|}, & \text{and } \vec{o}_j \vec{s}_i \cdot \vec{r}_{\phi_j} - \|\vec{o}_j \vec{s}_i\| \cos(\frac{A_o}{2}) \geq 0, \\ 0, & \text{otherwise} \end{cases}$$



Wireless Charger Deployment with Communication Constraint, Fig. 1 Directional charging model

where α and β are two constant parameters which are determined by the surrounding environment (Dai et al. 2016a, 2017); $\|\vec{s}_i \vec{o}_j\|$ denotes the distance between s_i and o_j .

Moreover, when a device o_j receives power from multiple wireless chargers, the received power of o_j is from all its ambient chargers power set (Dai et al. 2016a, 2017).

Problem Formulation

Because of hardware constraints, the received power of a rechargeable device is limited to some threshold P_{th} , so the *charging utility* of a device o_j can be represented as follows.

$$U(x) = \begin{cases} c_u \cdot x, & x \leq P_{th} \\ c_u \cdot P_{th}, & x > P_{th} \end{cases} \quad (1)$$

where Eq. (1) shows that the charging utility of the device o_j is first proportional to the total received power and then becomes a constant when the total received power exceeds the threshold P_{th} . The problem of placing B directional wireless chargers toward maximizing the overall charging utility of the N rechargeable devices under the constraint that all the deployed chargers should be connected is considered. A *strategy* of a charger is defined as a two-tuple $\langle p_i, \theta_i \rangle$, where p_i is the candidate position and θ_i is the orientation angle of the charger placed at the candidate position p_i . For the same candidate position p_i , which is allowed to place at most l_{p_i} chargers, say, *budget constraint* of the candidate position. Γ_{p_i} is used to represent the set of the strategies for candidate position p_i , $\Gamma_{p_i} = \{\langle p_i, \theta_{1i} \rangle, \dots, \langle p_i, \theta_{l_{p_i}i} \rangle\}$. As there are at most B chargers, $\Gamma = \bigcup_{i=1}^M \Gamma_{p_i}$ denotes the set of strategy. With all the above definitions, the considered Connected wIReless Charger pLacement problem (**CIRCLE**) can be formulated as follows.

$$\begin{aligned} (\mathbf{P1}) \quad & \max \sum_{j=1}^N U \left(\sum_{\langle s_i, \theta_i \rangle \in A} \Pr(s_i, \theta_i, o_j, \phi_j) \right) \\ & s.t. \quad A \in F^\alpha \quad \text{and} \quad \text{Subgraph } A \text{ is connected} \end{aligned} \quad (2)$$

Approximation Algorithm for CIRCLE-R

As CIRCLE is NP-hard and difficult to be directly tackled, in this section, a relaxed version of CIRCLE (CIRCLE-R for short) is first to be considered, that is, choosing at most B connected candidate positions at which there should place one single omnidirectional charger to maximize the overall charging utility. First, the proof that CIRCLE-R is a problem of maximizing a submodular set function with connectivity constraint is given. Then, a constant approximation ratio algorithm that outperforms the state-of-the-art algorithm by at least 1.5 times in terms of achieved approximation ratio is designed.

Problem Formulation for CIRCLE-R

Recall that CIRCLE-R is to find B connected candidate positions for placing B omnidirectional chargers with optimal charging utility. Therefore, CIRCLE-R can be defined as follows.

$$\begin{aligned}
 (\text{P2}) \quad & \max \sum_{j=1}^N U \left(\sum_{s_i \in A'} \Pr(s_i, 2\pi, o_j, \phi_j) \right) \\
 \text{s.t.} \quad & A' \in \mathcal{P} \text{ and Subgraph } A' \text{ is} \\
 & \text{connected}
 \end{aligned} \tag{3}$$

where $f'(A') = \sum_{j=1}^N U \left(\sum_{s_i \in A'} \Pr(s_i, 2\pi, o_j, \phi_j) \right)$ is the charging utility for B omnidirectional chargers.

Some notions are presented to facilitate further analysis.

Definition 1 (Kuo et al. 2015) Given a finite set V , a real-valued function f on the set of subsets of V is called a submodular set function if and only if it satisfies one of the following two equivalent conditions.

1. $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$, $\forall A, B \subseteq V$;
2. $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$, $\forall A \subseteq B \subseteq V$ and $v \in V \setminus B$.

With these definitions, the critical lemma is presented as follows.

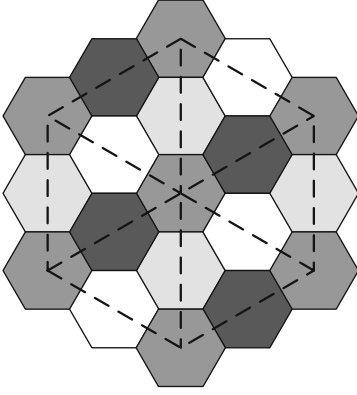
Lemma 1 *Given a subset A' of the candidate position collection $\mathcal{P}$, define $f'(A') = \sum_{j=1}^N U \left(\sum_{s_i \in A'} \Pr(s_i, 2\pi, o_j, \phi_j) \right)$, then $f'(A')$ is a submodular set function.*

Solution for CIRCLE-R

As Lemma 1 shows, the approximation algorithm described in Algorithm 2 for CIRCLE-R that consists of four steps is designed. First, the area is divided into uniform regular hexagon cells and classifies them into different groups, which is called *region division* (Steps 1–2). Second, the profit of each candidate position in the same cell in the same group is set as its caused increment to the overall charging utility using a greedy algorithm, while zero profit to candidate positions in the remaining group is assigned (Step 4). This step is called as *profit assignment*. Third, the problem of finding a connected subgraph with maximum charging utility is converted to the problem of finding a Quota Steiner Tree (QST), called as *finding QST* (Steps 6–13). Finally, as the obtained QST may contain more than B vertices, a dynamic programming is used to find the best subtree with at most B vertices in the QST, called as *calculating the best subtree* (Step 14). Next, the detail for these four steps is introduced in the following four subsections.

Region Division

A region H is found in the Euclidean plane such that not only it is large enough to cover all the N devices and the M candidate positions but it also can be partitioned into uniform smaller regular hexagonal cells with side length of $l = \frac{1}{2}R_c$. As such, the chargers in a same small regular hexagonal cell can communicate with each other. To make no device be charged by the different cells in the same group, the number of interval cells of different cells in the same group should be at least I along the division line as the dashed lines as Fig. 2 shows, where $I = \left\lceil \frac{2D}{\sqrt{3}l} \right\rceil$. Then these cells are partitioned into K different groups



Wireless Charger Deployment with Communication Constraint, Fig. 2 An instance of region division

where $K = I^2 + 2I + 1$ such that in each group, the number of interval cells between adjacent cells is exactly I . Without loss of generality, R_c is set to be one to simplify the derivation process. For example, as shown in Fig. 2, $I = 1$ is obtained given that $D = \frac{\sqrt{3}}{4}R_c$ and obtain four groups as shown in four different colors. In particular, let P_a represent the candidate position set in the same group where $a \in \{1, \dots, K\}$ and cl_i represent the candidate position set in the i -th regular hexagonal cell.

Profit Assignment

At this step, a greedy algorithm is used to assign a profit $W^* : \mathcal{P} \rightarrow \mathbb{Z}_0^+$ for the candidate position set $\mathcal{P}$ as described in Algorithm 1. For all candidate positions in a fixed cell of some group, a vertex p^* is greedily picked that has the maximum marginal charging utility that is rounded down to an integer. Note that the rounding-down process is necessary for assigning profit at the next step of finding a QST. With each selected vertex, its profit $w^*(p^*)$ is set to be the gained marginal charging utility. This step is repeated until all the candidate positions of the same group have picked and, then, set the profit for the other candidate positions of other groups to zero. Note that the order of choosing cells in the same group for assigning profit does not matter, because any pair of chargers in two different cells in the same group would never charge the same device.

Algorithm 1: Greedy algorithm for profit assignment

Input: Candidate position set $\mathcal{P}$, device set O , cell group P_a

Output: Profit assignment function $W^* : \mathcal{P} \rightarrow \mathbb{Z}_0^+$

```

1 for All  $p \in \mathcal{P}$  do
2    $w^*(p) = 0$ ;
3 for All  $cl_i \in P_a$  do
4    $X \leftarrow \emptyset$ ;
5   while  $|cl_i| \neq 0$  do
6      $p^* \leftarrow$ 
7        $\arg \max_{p \in cl_i \setminus X} [f'(X \cup \{p\}) - f'(X)]$ ;
8        $w^*(p^*) \leftarrow [f'(X \cup \{p^*\}) - f'(X)]$ ;
9      $X \leftarrow X \cup \{p^*\}$ ;
9 Output the profit value  $w^*$  for each candidate position.
```

Finding Quota Steiner Tree

After calculating the new profit for all the candidate positions, this profit is used for running a Quota Steiner Tree algorithm to find a tree T^* of size at most $4B$ whose quota is at least $\frac{1}{K}(1 - \frac{1}{e})OPT$, where OPT is the optimal profit. First, the definition of the Quota Steiner Tree is presented.

Definition 2 Quota Steiner Tree (QST). Given an undirected graph $G = (V, E)$, a profit function $p : V \rightarrow \mathbb{Z}_0^+$ on the vertices, a cost function $c : E \rightarrow \mathbb{Z}_0^+$ on the edges, and an integer (quota) q . A Quota Steiner Tree is the subtree T that minimizes $\sum_{e \in E(T)} c(e)$, subject to $\sum_{v \in V(T)} p(v) \geq q$.

The Quota Steiner Tree algorithm based on Johnson et al. (2000) and Garg (2005) achieves 2-approximation ratio. Therefore, if there exists a tree with at most $2B$ vertices and its quota being at least $\frac{1}{K}(1 - \frac{1}{e})OPT$, then the algorithm based on Johnson et al. (2000) and Garg (2005) can be used to find a tree T^* of size at most $4B$ with the quota of at least $\frac{1}{K}(1 - \frac{1}{e})OPT$.

Calculating the Best Subtree

At the last step, a tree T^* of size at most $4B$ is found. To obtain a feasible tree $\bar{T}$ of size at most B , there should calculate the best subtree

$\bar{T}$ of size at most B which has the highest total profit. A dynamic programming is used, which can be solved in polynomial time, to calculate the best subtree. Formally, the following problem is given. Given a tree $T^* = (V^*, E^*)$ of $4B$ vertices, each vertex has a profit $W^* : V^* \rightarrow \mathbb{Z}_0^+$, and an integer B , find a subtree $\bar{T}$ of size B which maximizes the total profit $w^*(\bar{T}) = \sum_{v \in \bar{T}} w^*(v)$, this subtree $\bar{T}$ is called the best subtree of T^* . Let the tree T^* be rooted at an arbitrary vertex and T_v^* denote the subtree of T^* rooted at a vertex v . Some notations used in the recurrence formula for the dynamic programming are introduced. $F(v, i)$ represents the best solution of at most i vertices completely contained inside T_v^* . $G(v, i)$ represents the best solution of at most i vertices completely contained inside T_v^* such that v is a part of the solution. Thus, calculating the best subtree becomes calculating $F(\text{root}, B)$. Suppose the root has m children represented as $v_1, \dots, v_m$, the following recurrence formula is obtained.

$$\begin{aligned}
 F(\text{root}, B) &= \max\{ \max_{1 \leq j \leq m} \{G(v_j, B)\}, \\
 &\quad G(\text{root}, B)\}, \\
 G(\text{root}, B) &= w^*(\text{root}) \\
 &\quad + \sum_{z=1}^m \max_{t_z=B-1} \{ \sum_{1 \leq j \leq m} G(v_j, i_j) \},
 \end{aligned} \tag{4}$$

The expression of $F(\text{root}, B)$ indicates that the best subtree with size at most B may be contained in the subtree with root root or in the subtree with root being one of the children of root . In addition, $G(\text{root}, B)$ is the sum of the profit of the root and the optimal profit of the best subtree with size of $B - 1$ contained in root 's children vertices. By tracking the intermediate state of the dynamic program, the optimal candidate positions for placing chargers can be obtained.

Theoretical Analysis for CIRCLE-R

In this subsection, theoretical analysis of the approximation ratio for our algorithm of CIRCLE-R is presented.

Theorem 1 *Algorithm 2 achieves a constant approximation ratio of $\frac{1}{67.50D^2 + 58.46D + 12.66}$ where D is the charging radius of chargers. When $D \rightarrow \infty$, the approximation ratio approaches $\frac{1}{67.50D^2}$.*

Proof. The detailed proof is omitted here to save space. $\square$

Note that the algorithm proposed by Huang et al. (2015) for solving the budgeted connected sensor cover problem achieves $\frac{1}{12.66(8C^2 + 4\sqrt{2}C + 1)}$ - approximation ratio where C is the sensing radius of sensor, a counterpart of D in the considered scenario. This is the best result, for solving the problem of maximizing a submodular set function with connectivity constraint. Apparently, the achieved approximation ratio of the algorithm

Algorithm 2: Algorithm for choosing the placement positions

Input: Candidate position set $\mathcal{P}$, device set O , budget B
Output: Candidate position set T for placement

- 1 Use the same size of regular hexagon to divide plane H ;
- 2 Calculate the group number K , determine which group each candidate position belongs to;
- 3 **for each group** P_a **do**
- 4 Call the greedy algorithm described in Algorithm 1 to assign the profit to each candidate position;
- 5 $\text{left} \leftarrow 0, \text{right} \leftarrow w^*(\mathcal{P})$;
- 6 **while** $\text{left} + 1 \neq \text{right}$ **do**
- 7 $\text{OPT}_{\text{GUESS}} \leftarrow \text{left} + (\text{right} - \text{left})/2$;
- 8 Call the 2-approximation algorithm based on Johnson et al. (2000) and Garg (2005) for the QST problem with the instance $(\mathcal{P}, w^*)$ as the input to obtain a tree T^* with quota at least $\text{OPT}_{\text{GUESS}}$;
- 9 **if** $|T^*| \leq 4B$ **then**
- 10 $T^* \leftarrow T^*$;
- 11 $\text{left} \leftarrow \text{OPT}_{\text{GUESS}}$;
- 12 **else**
- 13 $\text{right} \leftarrow \text{OPT}_{\text{GUESS}}$;
- 14 Use a dynamic program to find the best subtree $\bar{T}$ of the tree T^* ;
- 15 Output the candidate position set T corresponding to the vertices in the best subtree $\bar{T}$.

improves at least 1.5 times compared with the algorithm in Huang et al. (2015).

Approximation Algorithm for CIRCLE

In this section, based on the analysis to CIRCLE-R, the original CIRCLE problem is considered that takes into consideration the directional charging pattern of chargers and the budget constraint for each candidate position (that is, each candidate position is allowed to place more than one chargers while not exceeding a given budget) compared with CIRCLE-R.

Problem Formulation for CIRCLE Without Budget Constraint for Each Candidate Position

After choosing the best B candidate positions for placing chargers in CIRCLE-R, CIRCLE problem becomes choosing the best orientation of each charger to maximize the charging utility. Therefore, CIRCLE problem can be defined as follows.

$$\begin{aligned}
 (\text{P3}) \quad & \max \sum_{j=1}^N U \left(\sum_{s_i \in T} \Pr(s_i, \theta_i, o_j, \phi_j) \right) \\
 \text{s.t.} \quad & \theta_i \in [0, 2\pi)
 \end{aligned} \tag{5}$$

where $f^*(T, \theta_i) = \sum_{j=1}^N U \left(\sum_{s_i \in T} \Pr(s_i, \theta_i, o_j, \phi_j) \right)$ is the overall charging utility considering the orientations of chargers.

Solution for CIRCLE Without Budget Constraint for Each Candidate Position

Candidate Orientation Angle Extraction

The following definition is given to facilitate further analysis.

Definition 3 (dominant coverage set) Given a set of devices $\mathcal{O}_i^1$ covered by a charger s_i with some orientation, if there does not exist another set of devices $\mathcal{O}_i^2$ covered by the charger s_i with

Algorithm 3: Algorithm for candidate orientation angle extraction

Input: Charger position set T , device set O , charging angle A_s

Output: Collection of sets of charger orientation angles A^θ

```

1 for all  $T_i \in T$  do
2   Find the subset of devices  $O_i \subseteq O$  that satisfy
    $Distance(o_i, T_i) \leq D$  for each device  $o_i$  in set  $O_i$ ;
3   Rotate the charger with charging angle  $A_s$ 
   anticlockwise to cover the devices in  $O_i$  one by
   one until there is some covered device going to
   be uncovered. During the rotating process, if
   the rotated angle is larger than  $2\pi$ , then
   terminate;
4   Add the current covered set of devices to the
   collection of dominant coverage sets and add
   the current orientation angle  $\theta_i^1$  to the set  $A_i^\theta$ ;
5   Rotate the charger anticlockwise until a new
   device in  $O_i$  is included in the covered set. If
   the rotated angle is larger than  $2\pi$ , then
   terminate. If not, goto Line 3;
6    $A^\theta \leftarrow A^\theta \cup \{A_i^\theta\}$ ;
7 Output the collection of sets  $A^\theta$ .
```

some other orientation such that $\mathcal{O}_i^1 \subset \mathcal{O}_i^2$, then $\mathcal{O}_i^1$ is a dominant coverage set.

As any possible covered set of devices is a subset of a certain dominant coverage set, rather than considering all possible orientation angles of a charger in a candidate position, there only need to consider those orientation angles, called candidate orientation angles, that cover the dominant coverage sets of devices. The algorithm for candidate orientation angle extraction is described in Algorithm 3.

Selecting Orientation Angle

After candidate orientation angle extraction, a possible orientation angle set A_i^θ for each charger T_i is obtained. At this step, a greedy algorithm is used to find the best orientation angle θ_i for charger s_i that maximizes the increased charging utility when adding T_i to the charger set $\{T_1 \cup \dots \cup T_{i-1}\}$. The algorithm is described in Algorithm 4. For each charger, Algorithm 4 greedily adds the orientation θ_i to set θ such that the increment of the function f^* is maximized.

Algorithm 4: Greedy algorithm for selecting best orientation angle for each charger

Input: Charger position set T , device set O , charging angle A_s

Output: Orientation angle set θ for all chargers

- 1 Call Algorithm 3 with the instance (T, O, A_s) as the input to obtain the collection of sets of charger orientation angles A_i^θ ;
- 2 $\theta \leftarrow \emptyset$; $X \leftarrow \emptyset$;
- 3 **for all** $T_i \in T$ **do**
- 4 $\theta_i \leftarrow \arg \max_{A_{ij}^\theta \in A_i^\theta} f^*(X \cup \{T_i\}, A_{ij}^\theta)$
 $- f^*(X, \theta_{i-1})$;
- 5 $X \leftarrow X \cup \{T_i\}$;
- 6 $\theta \leftarrow \theta \cup \{\theta_i\}$;
- 7 **Output** best orientation angle set θ .

Algorithm 5: General algorithm for CIRCLE

Input: Candidate position set $\mathcal{P}$, device set O , budget B , and charging angle A_s

Output: Charger position set $\mathcal{T}$ and best orientation angle set Θ^*

- 1 For each candidate position p_i with budget constraint l_{p_i} , replace it with l_{p_i} candidate positions each of which is allowed to place only one charger, and thus construct a new candidate position set $\mathbb{P}$;
- 2 Call Algorithm 2 with the instance $(\mathbb{P}, O, B)$ as the input to choose the placement position set $\mathcal{T}$ of size at most B ;
- 3 Call Algorithm 4 with the instance $(\mathcal{T}, O, A_s)$ as the input to obtain the best orientation angle set Θ^* ;
- 4 **Output** $\mathcal{T}$ and Θ^* .

Solution for CIRCLE and Theoretical Analysis
Problem Redefinition

In this subsection, the more general situation is considered. Before describing the algorithm, CIRCLE problem is redefined as follows.

$$\begin{aligned}
 (\mathbf{P4}) \quad & \max \sum_{j=1}^N U \left(\sum_{s_i \in T} \sum_{\theta_i \in \Theta_i} \right. \\
 & \left. \Pr(s_i, \theta_i, o_j, \phi_j) \right) \\
 \text{s.t.} \quad & \theta_i \in [0, 2\pi) \quad \text{and} \quad \Theta_i \subseteq A_i^\theta \quad (6)
 \end{aligned}$$

Solution for CIRCLE

The general algorithm is illustrated in Algorithm 5. First, l_{p_i} same positions are generated

for each candidate position p_i in the candidate position set; the newly generated set is denoted as new candidate position set $\mathbb{P}$ (step 1). Second, Algorithm 2 is used on the new candidate position set $\mathbb{P}$ to choose the placement position (Step 2). Third, Algorithm 4 is used to select the best orientation angle for each placement position (Step 3).

Theoretical Analysis for CIRCLE

Theoretical analysis of the approximation ratio for the proposed algorithm described in Algorithm 5 for CIRCLE is given in this subsection.

Theorem 2 *Algorithm 5 achieves a constant approximation ratio of $\frac{A_s}{2\pi(67.50D^2 + 58.46D + 12.66)}$.*

Key Applications

Wireless Rechargeable Sensor Networks, Wireless Identification Sensing Platform, Millimeter Wave Cellular Networks, etc.

References

- Cheng Z et al (2015) Design and loss analysis of loosely coupled transformer for an underwater high-power inductive power transfer system. *IEEE Trans Magn* 51(7):1–10
- Chiu TC et al (2012) Mobility-aware charger deployment for wireless rechargeable sensor networks. In: *IEEE APNOMS*, pp 1–7
- Dai H et al (2016a) Omnidirectional chargeability with directional antennas. In: *IEEE ICNP*, pp 1–10
- Dai H et al (2016b) Radiation constrained wireless charger placement. In: *IEEE INFOCOM*, pp 1–9
- Dai H et al (2017) Optimizing wireless charger placement for directional charging. In: *IEEE INFOCOM*, pp 1–9
- Fusco G et al (2009) Selection and orientation of directional sensors for coverage maximization. In: *IEEE SECON*, pp 1–9
- Garg N (2005) Saving an epsilon: a 2-approximation for the k-MST problem in graphs. In: *Proceedings of the thirty-seventh annual ACM symposium on theory of computing*, ACM, pp 396–402
- Han X et al (2008) Deploying directional sensor networks with guaranteed connectivity and coverage. In: *IEEE SECON*, pp 153–160
- He S et al (2013) Energy provisioning in wireless rechargeable sensor networks. *IEEE Trans Mob Comput* 12(10):1931–1942

- Khorster E et al (2006) Approximating optimal visual sensor placement. In: 2006 IEEE international conference on multimedia and expo. IEEE, pp 1257–1260
- Huang L et al (2015) Approximation algorithms for the connected sensor cover problem. In: International computing and combinatorics conference. Springer, pp 183–196
- Johnson DS et al (2000) The prize collecting Steiner tree problem: theory and practice. In: SODA, vol 1, p 4
- Khuller S et al (2014) Analyzing the optimal neighborhood: algorithms for budgeted and partial connected dominating set problems. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms, society for industrial and applied mathematics, pp 1702–1713
- Kuo TW et al (2015) Maximizing submodular set function with connectivity constraint: theory and application to networks. *IEEE/ACM Trans Netw* 23(2): 533–546
- Lu X et al (2016) Wireless charging technologies: fundamentals, standards, and network applications. *IEEE Commun Surv Tutor* 18(2):1413–1452
- Madhja A et al (2015) Distributed wireless power transfer in sensor networks with multiple mobile chargers. *Comput Netw* 80:89–108
- Mahmoudi A et al (2014) Design, analysis, and prototyping of a novel-structured solid-rotor-ringed line-start axial-flux permanent-magnet motor. *IEEE Trans Ind Electron* 61(4):1722–1734
- Mascareñas D et al (2008) Wireless sensor technologies for monitoring civil structures. *Sound Vib* 42(4): 16–21
- RamRakhyani AK et al (2011) Design and optimization of resonance-based efficient wireless power delivery systems for biomedical implants. *IEEE Trans Biomed Circuits Syst* 5(1):48–63
- Yu N, Dai H, Liu A, Tian B (2018) Placement of connected wireless chargers. In: Proceedings of the IEEE INFOCOM, pp 387–395

Wireless Charger Placement with Communication Constraint

- [Wireless Charger Deployment with Communication Constraint](#)

Wireless Communication

- [Wireless Fading Channel Model for 5G and Future](#)

Wireless Communication in Data Center

Shiwen Liu¹ and Kan Zheng²

¹Intelligent Computing and Communication Lab, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

²Intelligent Computing and Communications (IC²) Lab, Wireless Signal Processing and Networks (WSPN) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Synonyms

[Wireless data centers](#); [Wireless technology in data centers](#)

Definitions

Wireless communication is the information transfer between two or more points that are not connected by an electrical conductor. The information is transmitted via the propagation of electromagnetic wave signals in free space. Wireless communication technologies can be incorporated in data centers (DCs) to overcome the cabling complexity and hotspots problems encountered by conventional wired data centers.

Historical Background

A data center is a physical or virtual centralized repository for storing, managing, and disseminating data. The concept of DC network (DCN) refers to the network infrastructure that provides intra- and inter-DC communications, which has received much attention in network research field. Generally, when designing a DCN, the designers need to consider several factors including scalability, latency, reliability, cost, and other practical details (Mudigonda et al. 2011).

Most existing DCNs rely on wired networking technologies, i.e., the communication service is implemented by using copper cables or fiber optics. The design and implement of wired DCNs face some inevitable problems, which lead to unsatisfactory network performance. Since the traffic patterns of DCN applications are difficult to predict, the conventional tree-based topology in wired DCNs often encounters the hotspot problem and flow congestion, especially when a burst of traffic occurs. In order to adapt to as many traffic patterns as possible, the cables and switches have to be over-deployed at high cost when designing the network (Huang et al. 2013). Moreover, when adding new servers to the network, considerable rewiring of the existing devices may be required, which increases the cabling complexity. The cable bundles behind the racks also cause problems of inefficient cooling and high maintenance cost. Although researchers have been working on new architectures (e.g., fat tree, DCell, BCube) in wired DCNs, the problems of hotspot and cabling complexity remain unsolved.

Foundations

Due to the limitations and problems in the existing wired DCNs, researchers have been working on incorporating wireless communication technologies in DCNs. Up to now, 60 GHz radio frequency (RF) and free space optics (FSO) are two candidate technologies which are mostly discussed.

60 GHz RF Technology

RF communication operates in the band of millimeter wave (mmWave) and benefits from the high frequency and large spectrum. Ramachandran et al. first introduce the idea of utilizing 60 GHz technology in DCNs (Ramachandran et al. 2008). It is pointed out that the radios in the license-free 60 GHz range give them operational advantages over others for providing link connectivity in data centers. It is suggested that building indirect LOS links by using reflectors for intra-rack communications,

and creating LOS, indirect LOS of multi-hop links for inter-rack communications.

In order to enhance the performance of conventional DCNs, flyways are designed to add additional on-demand links to wired DCNs (Kandula et al. 2009). The flyways can be wired switches or wireless transceivers. The use of wireless flyways can reduce the completion time of the demands and hence improve the performance of the network. Adopting 60 GHz wireless technology in wireless flyways can speed up DCN applications to some extent (Halperin et al. 2011).

When it comes to the topology design, it is possible to incorporate 60 GHz technology in the typical DCN topologies, i.e., hierarchical and fat-tree architectures, and implement direct LOS wireless links among the servers and switches (Prakash 2014). In addition to emulating well-known topologies, some researchers also work on designing new topologies for wireless DCNs. A novel DCN topology can be modeled as a mesh of Cayley graphs (CG). A semi-regular mesh topology is implemented by incorporating 60 GHz wireless transceivers and switching logic within each server node and arranging them in cylindrical racks. In this novel design, both the physical and the logical topologies are different from the typical wired topologies. Moreover, Cayley DCN can perform better in packet delivery latency and failure tolerance when compared to fat-tree DCN (Shin et al. 2012).

FSO Technology

FSO communication, also known as optical wireless communication (OWC), combines the advantages of both wireless and optical communication systems. FSO modulates light beam which propagates wirelessly from one point to another [8]. Light-emitting diodes (LEDs) and laser diodes can act as the light source at the transmitter in an FSO link.

The design and applications of power smart FSO links in DCNs have been widely discussed. In order to improve the mobility and modularity, containerized DCNs can be used where servers and other devices are arranged in a container. FireFly offers an inter-rack network solution for

wireless DCN design. The FSO transceivers are arranged on the top of racks (ToRs). Moreover, all FSO links are reconfigurable by using switchable mirrors (SMs) or Galvo mirrors (GMs). SMs can be switched to mirror or transparent states, and GMs can be rotated within a limited a range (Hamedazimi et al. 2014). In this way, the light beam in an FSO link can be deflected in the direction as designed. However, the links are limited to unicast. A new switching element, namely, tri-state switching element (T-SE), has three optional states are reflective, transmissive, and splitting state. When a T-SE is in the splitting state, the incident beam can be split into several copies and hence it realizes multicast (Hamza et al. 2016).

Challenges in Wireless DCNs

As mentioned above, the design of a robust DCN has to meet some requirements, e.g., network capacities, data transmission rate, interference, cooling efficiency, power consumption, maintenance cost, etc. The utilization of wireless technologies in DCNs also faces some challenges in meeting the requirements. For 60 GHz technology, the practical bandwidth is lower. Appropriate bandwidth allocation and interferences alleviation are required. In wireless DCN utilizing FSO technology, as all the point-to-point FSO links need accurate alignment, the choice of light source and the vibration of the light beam have to be carefully considered (Turner 2010). Moreover, using artificial light for the illumination of DCs may cause interference of the FSO links.

Key Applications

Wireless technologies can be incorporated in data centers to build wireless DCNs which can overcome the limitation of wired DCNs.

Cross-References

- [Data Center Network Virtualization](#)
- [Wireless Network](#)

References

- Halperin D, Kandula S, Padhye J, Bahl P, Wetherall D (2011) Augmenting data center networks with multi-gigabit wireless links. In: ACM sigcomm conference, pp 38–49
- Hamedazimi N, Qazi Z, Gupta H, Sekar V, Das SR, Longtin JP, Shah H, Tanwer A (2014) Firefly: a reconfigurable wireless data center fabric using free-space optics. *ACM sigcomm Comput Commun Rev* 44(4):319–330
- Hamza AS, Deogun JS, Alexander DR (2016) Rearrangeable non-blocking multicast FSO switch using fixed switching elements. In: IEEE global communications conference, pp 1–6
- Huang H, Liao X, Li S, Peng S, Liu X, Lin B (2013) The architecture and traffic management of wireless collaborated hybrid data center network. In: ACM sigcomm conference on sigcomm, pp 511–512
- Kandula S, Padhye J, Bahl V (2009) Flyways to decongest data center networks. In: ACM workshop on hot topics in networks
- Mudigonda J, Yalagandula P, Mogul JC (2011) Taming the flying cable monster: a topology design and optimization framework for data-center networks. *Sci China Phys Mech Astron* 53(2):369–379
- Prakash R (2014) 60 ghz wireless links in data center networks. *Comput Netw Int J Comput Telecommun Netw* 58(2):192–205
- Ramachandran K, Kokku R, Mahindra R, Rangarajan S (2008) 60 ghz data-center networking: Wireless? Worry less? NEC Laboratories America
- Shin JY, Weatherspoon H, Kirovski D (2012) On the feasibility of completely wireless datacenters, pp 3–14
- Turner J (2010) Effects of data center vibration on compute system performance. In: Usenix conference on sustainable information technology, pp 5–5

Wireless Community Networks

- [Sharing Economics](#)

Wireless Credentials

- [Authentication](#)

Wireless Data Centers

- [Wireless Communication in Data Center](#)

Wireless Data Security

- [Distributed Storage Security](#)

Wireless Edge Caching

Zhiwen Hu, Zijie Zheng, and Lingyang Song
School of Electronics Engineering and Computer
Science, Peking University, Beijing, China

Synonyms

[Edge caching](#); [Mobile caching](#); [Mobile content delivery](#); [Wireless caching](#)

Definition

Wireless edge caching is the technique that distributes duplicate contents from the original Internet server to the cache-enabled devices in wireless networks. These devices often refer to the gateways in the core networks (CNs) or the base stations in the radio access networks (RANs), and in some cases, cache-enabled mobile devices (such as cell phones) can also be taken into account. With wireless edge caching, mobile users are able to obtain the desired contents directly from those nodes instead of tracing back to the original Internet server. Both the access delay of the mobile users and the backhaul traffic of the wireless networks can be reduced.

Historical Background

With the proliferation of the Internet in the late 1990s, popular web services started to face the

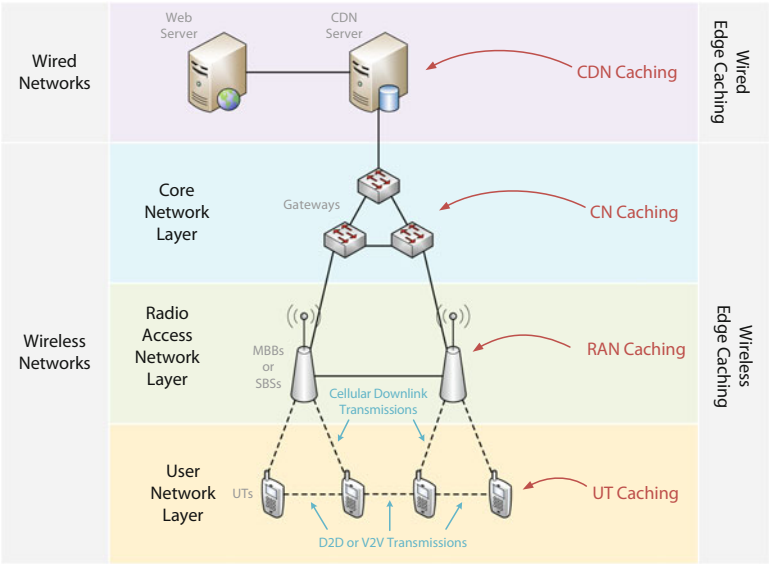
problem of congestion and bottlenecks due to frequent demands from users (Pathan and Buyya 2007). Content delivery network (CDN) was then proposed to deal with this problem by replicating the original contents and caching them into other servers with different geographical locations (Douglass and Kaashoek 2001). CDN cache servers are also called as edge servers, since they are usually closer to the end users who require the contents. Correspondingly, CDNs have constituted the most important part of *edge caching* since then.

The relevant concept, wireless edge caching, comes from the demand for rich multimedia services over wireless networks in recent years (Wang et al. 2014). Mobile users access the Internet through the RANs and the CNs, where the lengths of the intermediate links are longer than that of the links of traditional wired users. Thanks to the evolving 4G and 5G systems, the RANs and the CNs are expected to be endowed with the ability to cache contents (Ramanan et al. 2013). Along with the cache-enabled user devices, all the related nodes in the wireless networks can be seen as edge caching servers, which forms the concept of *wireless edge caching*.

Foundations

Caching is the technique that essentially reduces duplicated transmissions in the system with the cost of additional caching storages and computing resources. Such trade-off is profitable as long as the reduction of duplicated transmissions is valuable enough to counteract all the costs (Erman et al. 2011). Due to the recent development of storage techniques, the cost of deploying caching storages becomes more and more negligible. Based on the long-tail property of the distribution of popularity of contents, only a small proportion of popular contents are requested by a large portion of mobile users, which makes caching a promising way to deal with duplicated transmissions. From the view of the wireless network operator (WNO) who runs the devices of

Wireless Edge Caching,
Fig. 1 The basic
architecture of wireless
communication networks



the wireless network, wireless edge caching can avoid unnecessary usages of bandwidth-limited links in the wireless network and thus reduces the probability of congestion during rush hours. From the view of the mobile users who aim to download their desired contents, wireless edge caching is also able to averagely bring lower access latency and improve user experience.

The basic architecture of a wireless communication network has a multilevel topology, as illustrated in Fig. 1. To connect to the Internet, user terminals (UTs) in the user network (UN) layer have to rely on the CNs and the RANs, which contains the devices managed by the WNO. The CN mainly consists of different kinds of gateways, and the RAN mainly consists of macro base stations (MBBs) and small base stations (SBSs). A practical structure of a wireless network could be more complicated, but the basic idea remains the same. As CDNs can provide caching services for mobile users with the assistance of CDN servers, other network nodes in the CN, the RAN, and the UN layers can also be applied as caching storages to move the replicate contents closer to the edge of the wireless networks. Caching in the CN, the RAN, and the UN layers can be considered as the three major aspects of wireless edge caching. While CN caching has the advantage of large caching storages and high hit rate, RAN caching

leads to much lower latency and also saves the backhaul bandwidth of the RAN. With the help of device-to-device (D2D) and vehicle-to-vehicle (V2V) transmissions, UT caching brings the contents closest to end users. The system would be quite efficient by utilizing hybrid caching which includes all the caching layers in wireless edge caching along with wired edge caching.

Wireless edge caching strategies can be generally classified into two categories based on the knowledge of content popularity, which are *reactive caching* and *proactive caching*. Reactive caching refers to the caching schemes in which the cache-enabled devices are not aware of the popularity of the contents. Each device can only adjust its caching storage based on the local information (i.e., the history of the requests sent to this device) in a reactive manner. Exemplary reactive caching schemes include First In First Out (FIFO), Least Recently Used (LRU), and Least Frequently Used (LFU) (Lee et al. 2001). Proactive caching, however, refers to the caching schemes in which the cache-enabled devices are aware of the popularity of contents based on some centralized estimations and predictions. The contents are allocated into the caching storages by the WNO in a proactive manner. With enough estimation and prediction accuracy, proactive caching can achieve a better performance compared with reactive caching

(Baştuğ et al. 2014). Therefore, the recent rapid development of machine learning techniques is able to improve the efficiency of proactive caching. Another noticeable aspect of proactive caching is that the interest-relevant groups (e.g., WNOs, content providers, and mobile users) may influence the caching results with game interaction (Hu et al. 2016), which is not a vital concern in reactive caching.

From another perspective, wireless edge caching strategies can also be classified into *content-aware solutions* and *non-content-aware solutions*. In content-aware solutions, each content has a unique identity to distinguish from each other in the caching storages. Each caching device is aware of whether a specific content is cached in its storage. In non-content-aware solutions, caching devices cache the contents based on chunks, packets, flows, or even bytes. Contents are no longer stored as whole entities in the storages of those devices. Current deployed CN caching solutions include both content-aware caching and non-content-aware caching (Woo et al. 2013). Since the method being applied is very much alike with that of the traditional CDN caching, a network with CN caching is usually called as a mobile content delivery network (mobile CDN). RAN caching, however, is not ready to apply content-aware caching given the current hardware in the 4G system, because the RANs only establish tunnels between the UTs and the CNs and are not content-aware. Nevertheless, the concept of information-centric networking (ICN) architecture (Zhang et al. 2013) for the future Internet is a promising technique to enable content-aware RAN caching.

Key Applications

Wireless edge caching is a useful design consideration of establishing practical wireless networks.

Cross-References

- [Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks](#)

- [Mobile Content Distribution Architecture](#)
- [Mobile Edge Caching in HetNets](#)

References

- Baştuğ E, Bennis M, Debbah M (2014) Living on the edge: the role of proactive caching in 5G wireless networks. *IEEE Commun Mag* 52(8):82–89
- Douglis F, Kaashoek MF (2001) Scalable internet services. *IEEE Internet Comput* 5(4):36–37
- Erman J, Gerber A, Hajiaghayi MT, Pei D, Sen S, Spatscheck O (2011) To cache or not to cache: the 3G case. *IEEE Internet Comput* 15(2):27–34
- Hu Z, Zheng Z, Wang T, Song L, Li X (2016) Game theoretic approaches for wireless proactive caching. *IEEE Commun Mag* 54(8):37–43
- Lee D, Choi J, Kim JH, Noh SH, Min SL, Cho Y, Kim CS (2001) Lrfu: a spectrum of policies that subsumes the least recently used and least frequently used policies. *IEEE Trans Comput* 50(12):1352–1361
- Pathan AMK, Buyya R (2007) A taxonomy and survey of content delivery networks. *Grid Computing and Distributed Systems Laboratory. Technical Report 4*, University of Melbourne
- Ramanan BA, Drabeck LM, Haner M, Nithi N, Klein TE, Sawkar C (2013) Cacheability analysis of http traffic in an operational LTE network. In: *Wireless Telecommunications Symposium (WTS)*, 2013. IEEE, Phoenix, AZ, pp 1–8
- Wang X, Chen M, Taleb T, Ksentini A, Leung VCM (2014) Cache in the air: exploiting content caching and delivery techniques for 5G systems. *IEEE Commun Mag* 52(2):131–139
- Woo S, Jeong E, Park S, Lee J, Ihm S, Park K (2013) Comparison of caching strategies in modern cellular backhaul networks. In: *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, Taipei, Taiwan, pp 319–332
- Zhang G, Li Y, Lin T (2013) Caching in information centric networking: a survey. *Comput Netw* 57(16):3128–3141

Wireless Fading Channel Model for 5G and Future

Jianhua Zhang
Beijing University of Posts and
Telecommunications, Beijing, China

Synonyms

Channel impulse response (CIR); Wireless communication

Definition

The *communication channel* is a kind of physical medium or pathway such as air, water, and vacuum which can be used to carry information from a transmitter to its receiver. According to electromagnetic radiation theory, a *wireless channel* refers to the wave conveying a radio frequency (RF) signal between a source and its receiver in space. Thus, in wireless communication systems, the propagation characteristics of a *wireless channel* directly determine the performance limit of the wireless communication system. *Wireless channel model* is applying physical or mathematical methods to describe its fading characteristics and regenerate channel impulse response (CIR). While due to the complex reflection, diffraction, and scattering of the electromagnetic wave in real environments, an accurate and efficient channel model is crucial to the research of novel transmission techniques, the evaluation of standardization proposals, and the actual deployment of wireless system etc.

Historical Background

Since J. C. Maxwell forecasted the existence of the electromagnetic wave in 1865 and H. R. Hertz experimentally proved in 1887, wireless mobile communication system has experienced rapid development, especially, terrestrial mobile communication from the first generation (1G) in 1980s to the fourth generation (4G) around 2010, and fifth generation (5G) is expected around 2020. Wireless channel models have been widely studied over the world since then; a lot of research activities were carried out based on different requirements (Zhang 2012).

The initial research of wireless channel models can be traced back to 1940s, when Bell Labs gave the Ricean statistic model, which defines the amplitude fading distribution of a narrowband channel with a line of sight (LoS) path (Rice 1944). In the 1950s, Lincoln Laboratory proposed a time-frequency two-dimensional statistic model in order to describe

the resolvable multipath in delay domain. That is the early concept of tap-delay-line (TDL) model for second- (2G) and third-generation (3G) mobile communication system research. From 1970s, the field measurement campaigns were carried on by the sounder either by single pulse scheme or sequence correlation scheme (Cox 1977; Turin et al. 1972). Many classical empirical models appeared, for example, Egli, Okumura, Hata, Walfisch, and Bertoni models. In particular, some organizations and institutions have also defined their standardized channel models, i.e., Global System for Mobile communications (GSM) 05.05 and International Telecommunications Union-Radio (ITU-R) M. 1225 for 2G and 3G radio transmission techniques (RTT) research and development. GSM 05.05 is specified by European Telecommunication Standard Institute (ETSI) which covers Rural Area (6 taps), Hilly Terrain (12 taps), and Typical Urban (12 taps) cases (ETSI EN 300 910 2000). ITU-R M. 1225 is defined for IMT-2000 (3G) which includes TDL models for Indoor office, Outdoor to indoor, and Pedestrian, Vehicular cases (ITU-R M.1225 1997).

Foschini and Gans (1998) and Telatar (1999) discovered that multiple-input multiple-output (MIMO) can provide remarkable spectral efficiency; the channel model is naturally expanded to space-time-frequency three dimensions in order to support MIMO study. Thus, double-directional model is proposed which combines the angle of departure (AOD) from transmitter with the angle of arrival (AOA) to receiver (Molisch et al. 2006). Moreover, cluster is revolutionarily introduced into modeling in order to describe the multipath components (MPCs) more accurately with low complexity (Saleh and Valenzuela 1987). Based on the above progress, geometry-based stochastic channel model (GBSM) was born and has been adopted by standardization organizations, i.e., spatial channel model (SCM) of 3rd generation partnership program (3GPP) and ITU-R M.2135 for IMT-Advanced evaluation.

To meet 5G new requirements, including 3D MIMO, millimeter wave (MM), and high speed, GBSM model is continuously evolved into cover

elevation domain, i.e., 3D MIMO channel model (Zhang et al. 2017a) and standardized into 3GPP TR 36.873 (3GPP TR 36.873 V1.0.0 n.d.). New features for millimeter (mm) wave is studied (Rappaport et al. 2017; Zhang et al. 2017b) and standardized into 3GPP 38.900/901 (3GPP TR 38.900 2016; 3GPP TR 38.901 2017). ITU-R has also launched the draft group for IMT-2020 (5G) channel model since 2016 and its work is nearly finished (ITU-R M.2412 2017). Thus, 5G channel model is defined and issued after October 2017 in order to support the 5G system design and application before 2020.

Foundations

It is known that versatile objects exist in the real environments, like glass windows, concrete buildings, plants, moving cars, etc. The radio waves may be reflected, scattered or diffracted, arrive at the receiver through different paths as shown in Fig. 1. Thus the received signal is the sum of multiple radio waves with different phases and delays. Consequently, the transmitted signal experiences dramatic variation when going through the radio channel and such fading is plotted in Fig. 2 as the red line.

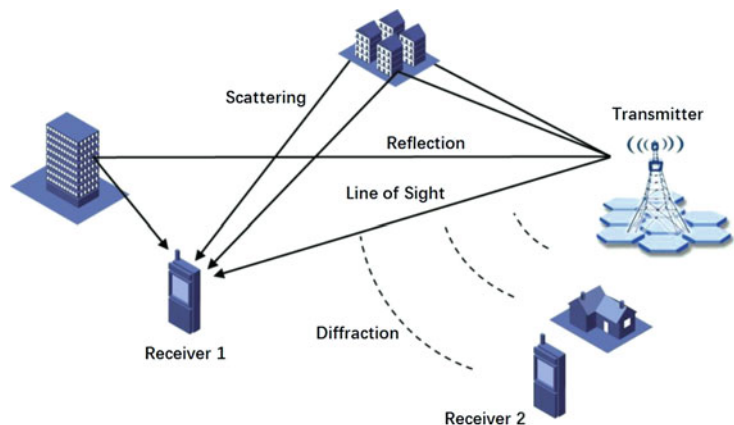
For simplicity, this fading is classified to *large-scale fading* and *small-scale fading*. Large-scale fading in Fig. 2 as blue line describes the average channel gain over a distance of tens to a few hundred wavelengths. It is important for coverage prediction and interference analysis

of a radio system. Typical large-scale fading is modeled into *Path loss* and *shadow fading*. Path loss in Fig. 2 as black line is the reduction in power density of a radio wave as it propagates through the distance. Shadow fading is caused by large terrain features between the transmitter and receiver. As for the small-scale fading, it describes the signal variation over a short distance scale, e.g., fractions of wavelengths. Small-scale fading determines the performance of air interface technologies. And the *small-scale fading* parameters, including delay spread, AoD and AoA, Doppler frequency, and complex magnitude of polarization, are acquired by first using a multidimensional joint parameter estimation algorithm from the measured data.

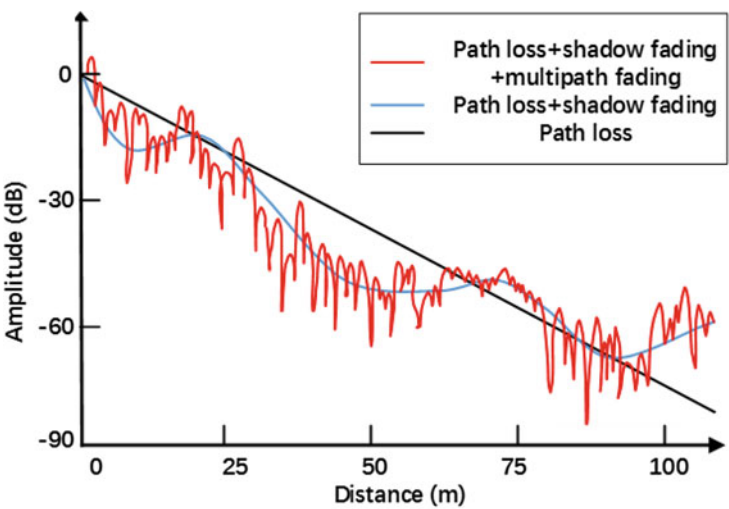
In order to describe the channel fading properties and regenerate CIR as red line in Fig. 2 accurately and efficiently, there are various channel model methodologies. Taking the MIMO channel model for example in Fig. 3, they can be classified into three categories, i.e., geometrical-based stochastic model, correlation-based model, and deterministic model. For geometry-based stochastic model, response of MPCs on the different elements of antenna array are geometrically calculated while delay spread, AoD and AoA, Doppler frequency, and complex magnitude of polarization are determined by empirical distribution. Because the model can reflect the real environment with low complexity, it is popular in standards, including ITU-R IMT-2020, WINNER+, 3GPP 36.873, 3GPP 38.900/901, and COST2100. The

Wireless Fading Channel Model for 5G and Future,

Fig. 1 The explanation of radio wave propagation in real environment



Wireless Fading Channel Model for 5G and Future,
Fig. 2 One example of large-scale fading and small-scale fading



Wireless Fading Channel Model for 5G and Future,
Fig. 3 The types of MIMO channel model

Stochastic model	Geometry based stochastic model	Correlation based model
	ITU-R IMT-2020 WINNER+ 3GPP 36.873 3GPP 38.900/901 COST 2100	Structural model Weichselberger model Kronecker model
Deterministic model	Ray-tracing method	
	Hybrid model	METIS map-based channel model

correlation-based channel model, especially Kronecker model, is well used in analyzing MIMO performance theoretically. Its CIR is expressed by channel correlation matrix which includes transmitting and receiving antennas' correlation parameters and channel variance. In terms of deterministic channel model, solution of Maxwell' s equations are used to determine the accurate CIR of a location by detailed geographic information. It is highly complex and relies on the precision of the geographic information. Recently, hybrid model and map-based model

from METIS are proposed for mm wave band channel model.

Key Applications

The growing research in the discussion of 5G wireless communications system has aroused the attention of the whole world since the successful commercialization of 4G. As 3GPP has preliminarily planned to open 5G new radio (NR) standardization process at the end of 2015;

it becomes urgent to establish related criterion for 5G NR evaluation. Compared with previous mobile communication systems like 3G or 4G, 5G and the beyond has been expected to include a bunch of new requirements, applications, and scenarios, e.g., massive and 3D MIMO, high frequency (millimeter wave), machine-to-machine, device-to-device, vehicle-to-vehicle, unmanned drone and high speed trains, satellite, etc.

Future Directions

Supporting the tremendously high data rates and constantly increasing new applications is the main task of 5G and future wireless communication system. To meet the huge demands, IMT-2020 is expected to offer the user experienced data rate to 100Mbps and peak data rate to 20Gbps. Furthermore, the vision of IMT-2020 presents the convergence of wireless communication, Internet, Internet of Things (IoT), and Machine-type Communication (MTC), which also brings increasing pressure to channel model research. Here remains some of the promising directions and open issues.

At present, mm wave and THz frequencies are now researched by major research institutions and industry. However, mm wave channels can be influenced by some fundamentally different propagation effects, such as atmospheric absorption, fog, rain, and shadows of human bodies and other material objects (Callum et al. 2017). Moreover, expensive measurement equipment and limited measurement data lead to that we can't master the comprehensive properties of mm wave and THz channel. Thus, we still need to put a lot of effort in order to explore so many bands of mm wave and THz, like cluster sparsity, spatial information, and suitable modeling methodology.

Due to the increased antenna number, larger bandwidth, and versatile complexity scenarios in 5G and future communication system, the measurement data reveals big volume feature. Thus, it is very natural that data mining schemes can be used to replace the conventional statistic schemes in channel characteristic analysis and some initial work already presented in (Ma et al.

2017). Moreover, artificial intelligent (AI) is also an intelligence exhibited by machines, which can find out the internal rules from data sets and then transform the rules to suitable models with learned parameters. By utilizing AI methods, we can cope with tasks such as feature extraction, classification, and prediction from the measurement data. Then we can apply them to analyze and reconstruct the real wireless channel more efficiently (Zhang 2016).

Wireless connectivity is an essential ingredient to Internet of Things (IoT), and 5G is a potential enabling technology bringing about spectrum, lower-cost connectivity to the IoT. But due to the rapidly changing propagation conditions that are typically found in IoT communication environments, the Doppler shift effects and the nonstationary characteristics of the wireless channel become exacerbated. Therefore, proper channel models are needed to provide insights into the physics of IoT radio reception.

References

- 3GPP TR 36.873 V1.0.0 (2014) 3rd generation partnership project; technical specification group radio access network; 3D channel model for LTE
- 3GPP TR 38.900 (2016) Study on channel model for frequency spectrum above 6 GHz
- 3GPP TR 38.901 (2017) Study on channel model for frequencies from 0.5 to 100 GHz
- Callum T, Mansoor S, Smith PJ, Dmochowski PA, Zhang JH (2017) Measurements and Models: massive MIMO for microwave and millimeter-wave frequency bands. *IEEE Trans Antennas Propag* 65:6505–6520
- Cox DC (1977) Multipath delay spread and path loss correlation for 910-MHz urban mobile radio propagation. *IEEE Trans on Veh Tech* 26:340–344
- ETSI EN 300 910 2000 Digital cellular telecommunications system (phase 2+): radio transmission and reception
- Foschini GJ, Gans MJ (1998) On limits of wireless communications in a fading environment when using multiple antennas. *Wirel Pers Commun* 6:311–335
- ITU-R M.1225 (1997) Guidelines for evaluation of radio transmission technologies for IMT-2000
- ITU-R M.2412 (2017) Guidelines for evaluation of radio interface technologies for IMT-2020
- Ma XC, Zhang JH, Zhang YX, Ma ZY, Zhang Y (2017) A PCA based wireless channel modeling method. In: *Proceeding of IEEE Inforcom, 5G & beyond workshop*, pp 1–6

- Molisch AF, Asplund H, Heddergott R, Steinbauer M, Zwick T (2006) The COST259 directional channel model-part I: overview and methodology. *IEEE Trans on Wirel Commun* 5:3421–3433
- Rappaport TS, Xing YC, MacCartney GR, Molisch AF, Mellios E, Zhang JH (2017) Overview of millimeter Wave communications for fifth-generation (5G) wireless networks-with a focus on propagation models. *IEEE Trans Antennas Propag* 65:6213–6230
- Rice SO (1944) Mathematical analysis of random noise. *Bell Sys Technical J* 23:282–332
- Saleh AA, Valenzuela R (1987) A statistical model for indoor multipath propagation. *IEEE J Sel Areas Commun* 5:128–137
- Telatar IE (1999) Capacity of multi-antenna Gaussian channels. *Eur Trans Telecommun* 10:585–595
- Turin GL, Clapp FD, Johnston TL, Fine SB, Lavry D (1972) A statistical model of urban multipath propagation. *IEEE Trans Veh Technol* 21:1–9
- Zhang JH (2012) Review on wideband MIMO channel measurement and modeling for IMT- advanced systems. *Chin Sci Bull* 57:1387–1400
- Zhang JH (2016) The interdisciplinary research of big data and wireless channel: a cluster-nuclei based channel model. *China Commun* 13:14–26
- Zhang JH, Zhang YX, Yu YW, Xu RJ, Zheng QF, Zhang P, MIMO 3D (2017a) How much does it meet our expectation observed from antenna channel measurements? *IEEE J Sel Areas Commun* 35(8):1887–1903
- Zhang JH, Tang P, Tian L, Hu ZX, Wang T, Wang HM (2017b) 6-100 GHz research progress and challenges for fifth generation (5G) and future wireless communication from channel perspective. *SCIENCE CHINA Inf Sciences* 60(8):1–16

Wireless Group Authentication

Anxi Wang
School of Computer and Software, Nanjing
University of Information Science and
Technology, Nanjing, China

Synonyms

Group authentication; Secret sharing

Definitions

Entities in wireless networks are always organized into group for network managers'

convenience. Group authentication is the entity identification strategy based on the identities or features of the wireless entities. Owing to the group structure, group authentication is very suitable for wireless networks. In addition, group authentication is one of the most important technologies for wireless communication in that network.

Historical Background

Group authentication is developed to provide better security for group-oriented network. In a wireless group-oriented network, there are multiple entities wanting to form a private network to exchange messages. In order to protect the exchanged messages, every entity in the network should authenticate others. According to the centralized authentication (CA) server's participation, group authentication schemes can be divided into three types: CA-centric type, entity-centric type, and CA-aided type.

In the CA-centric type, the CA takes the role of communicating with entities and authenticating entities. In 2006, Bhakti et al. (2007) proposed to adopt extensible authentication protocol (EAP) in the IEEE 802.1x Standard (Craig et al. 2002) for wireless ad hoc network. EAP has been thought to be the representation of CA-centric group authentication scheme.

In the entity-centric type, each entity authenticates group members. If there are n members in one group, each entity authenticates $(n - 1)$ times by applying conventional authentication scheme. The $O(n)$ complexity is the bottleneck for the widespread use of this type of group authentication schemes.

In 2013, Lein Harn (2013) proposed a secret sharing-based group authentication scheme, which authenticates all users at once. This group authentication scheme is based on an n -th order polynomial function, which was first applied to the field of cryptography by Shamir in 1979. This group authentication is named CA-aided type, which combines the advantages of the above two authentication methods. However, this group

authentication is unable to resist collusion attacks by more than n entities.

Many security and overhead issues have been found in these group authentication schemes. In the CA-centric type, network participants should fully trust the CA. The security and privacy may be leaked by the malicious CA or malicious attacker who attacks the CA. In the entity-centric type, the communication and computation cost to each entity is very huge. The heavy cost may badly affect the main functions of the wireless networks. In the CA-aided type, collusion attack is the main security risk. In addition, the secret shared among each entity takes up a lot of network bandwidth. How to overcome the security and overhead issues has become the top priority for promoting group authentication schemes.

Foundations

There are a series of group authentication methods in wireless networks. According to whether the protocol is based on encryption technology, it is divided into the following four types: key pre-distribution scheme, broadcast scheme, key update scheme, and secret sharing-based scheme.

Key pre-distribution-based scheme Many key distribution schemes are designed to provide secure communication among communication entities. Key pre-distribution scheme (Chan et al. 2003) is the most accepted key distribution method in wireless networks (Bechkit et al. 2013). The key generators provide secret keys to others before the network was formed. The key receivers are thought to be authenticated by the key generator. In wireless networks, the key generator and key receivers form a message exchange group. However, there are some drawback of the scheme, which cannot be ignored. Firstly, the group is fixed after it is generated. If the number of members in the group changes, the group authentication is also changed. Secondly, the authentication information increases linearly with increasing

number of members in the group. This is a major challenge to a user's storage and update capability. Thirdly, even if the group is accidentally attacked, the intragroup secret key cannot be changed.

Broadcast-based scheme Each entity sends a group key to a selected set of entities (Ren et al. 2009). In the broadcast scheme, which is based on a symmetric key, the sender has fixed authority to determine the set of receivers. In this case, the size of key has a minimum value (Perrig et al. 2005). In addition, safety is only guaranteed among a limited number of users. In the public key broadcast scheme, the key size problem can be avoided, but the threshold number of bad users is still set for safety. Note that the size of the ciphertext depends on the number of users, so it may be large in a large group. Therefore, the broadcast-based scheme is more flexible than key pre-distribution-based scheme.

Key update-based scheme This type is applicable to the case where the number of entities in the network changes during communication of the network (Wang et al. 2006). When an entity leaves the network, the network manager needs to rerun the authentication phase. In this method, the two questions must be considered. The operation is performed by the network manager, and the key is updated frequently. At the same time, whenever the number of entities changes, the authentication operation is repeated, which is unrealistic for actual wireless networks. Note that the key update-based schemes are stronger than the key pre-distribution-based schemes and the broadcast-based schemes.

Secret sharing-based scheme In addition to encryption certification, group authentication schemes have also been proposed based on different mathematical tools. After secret sharing schemes were first proposed by Shamir (1979), the scheme was applied to group authentication. In 2010, Harn (2013) proposed a basic t -secure m -user n -group authentication scheme($t; m; n$). The $(t; m; n)$ scheme can greatly reduce the user's computational cost. In 2015, Miao

et al. (Fuyou et al. 2015) found that the shared data in Harn's scheme is not bound to each other and introduced the notion of a randomized component (RC) to bind share data with all users and to protect the share data from being exposed to the outside without computational assumptions. Several researchers have designed new schemes to overcome the upper limit of t (Lin et al. 2012). However, the upper limit still exists in secret sharing schemes.

Key Application

Wireless group authentication scheme provides the secure communication among entities when they exchange message in wireless networks. In addition, when the authentication scheme accomplished, a group secret key can be generated for data protection (Ateniese et al. 1998; Lai et al. 2013). Group authentication schemes still should be improved to achieve high availability with high security.

Cross-References

- [Authentication](#)
- [Group Key Agreement for Wireless Networks](#)
- [Wireless Data Security](#)

References

- Ateniese G, Steiner M, Tsudik G (1998) Authenticated group key agreement and friends. In: Proceedings of the 5th ACM conference on computer and communications security. ACM, New York, NY, pp 17–26
- Bechkit W, Challal Y, Bouabdallah A, Tarokh V (2013) A highly scalable key pre-distribution scheme for wireless sensor networks. *IEEE Trans Wirel Commun* 12(2):948–959
- Bhakti MAC, Abdullah A, Jung LT (2007) Eap-based authentication for ad hoc network. In: Seminar Nasional Aplikasi Teknologi Informasi (SNATI)
- Chan H, Perrig A, Song D (2003) Random key pre-distribution schemes for sensor networks. In: Symposium on security and privacy. IEEE, Washington, DC, pp 197–213
- Craig JP et al (2002) 802.11, 802.1x, and wireless security. SANS Institute InfoSec Reading Room
- Fuyou M, Yan X, Xingfu W, Badawy M (2015) Randomized component and its application to (t, m, n) -group oriented secret sharing. *IEEE Trans Inf Forensics Secur* 10(5):889–899
- Harn L (2013) Group authentication. *IEEE Trans Comput* 62(9):1893–1898
- Lai C, Li H, Lu R, Shen XS (2013) SE-AKA: a secure and efficient group authentication and key agreement protocol for LTE networks. *Comput Netw* 57(17):3492–3510
- Lin HY, Yang CY, Hsieh MY (2012) Secure map reduce data transmission mechanism in cloud computing using threshold secret sharing scheme. In: Software engineering and knowledge engineering: theory and practice. Springer, Berlin/Heidelberg, pp 761–769
- Perrig A, Canetti R, Tygar JD, Song D (2005) The tesla broadcast authentication protocol. *Rsa Cryptobytes* 5
- Ren K, Yu S, Lou W, Zhang Y (2009) Multi-user broadcast authentication in wireless sensor networks. *IEEE Trans Veh Technol* 58(8):4554–4564
- Shamir A (1979) How to share a secret. *Commun ACM* 22(11):612–613
- Wang CL, Horng G, Chen YS, Hong TP (2006) An efficient key-update scheme for wireless sensor networks. In: International conference on computational science. Springer, Berlin/Heidelberg, pp 1026–1029

Wireless Jamming Attack

Ahmad Alagil and Yao Liu
Department of Computer Science and
Engineering, University of South Florida,
Tampa, FL, USA

Synonyms

[Anti-jamming](#); [Security](#); [Spread spectrum](#); [Wireless networks](#)

Definitions

In wireless networks, jamming attacks disrupt the legitimate communication between two nodes using interference signals. As a result, a channel between a sender and a receiver is blocked.

Historical Background

Over the last decades, wireless communications have been used widely to exchange data between multiple devices. In a wireless communication system, both a transmitter and a receiver share critical information during the transmission time. This makes the wireless channels to be subject to challenging aspects, such as noise and interference. Additionally, there are security concerns associated with wireless communication systems due to open nature of wireless networks, and the limitation of the power and computation capability of each device. Jamming attacks aim to target the channel of wireless communication systems and to prevent both a sender and a receiver from sharing data.

A jammer aims to prevent a legitimate sender and receiver from sending and receiving messages. He can simply emit noise signal to make the channel busy such that a sender is not able to transmit data. Also, a jammer can send junk data to a receiver. As a result, a receiver cannot recover the received message correctly. Both nonreactive and reactive jammers are considered as the most popular types of jamming attacks. A nonreactive jammer has no knowledge about the channel usage status. On the other hand, a reactive jammer senses the channel before it starts jamming. If the channel is idle, the jammer stays quiet. If a jammer detects transmission on the channel between a sender and a receiver, it starts emitting noise signal on the channel.

In fact, it is important to make the wireless communication between a sender and a receiver reliable by developing new schemes to mitigate the effects of jamming attacks. Frequency-hopping spread spectrum (FHSS) and direct-sequence spread spectrum (DSSS) are examples of spread spectrum techniques that have been widely used to defend against jamming attacks for wireless communication systems (e.g., Grover et al. 2014; Liu et al. 2010b). Traditional FHSS and DSSS assume that the sender and the receiver share the same secret key to encode messages. The shared secret key enables the communicators to generate the same frequency-hopping patterns

or spreading code sequences. Therefore, if a jammer knows the secret key, then he can jam the communication channel by using the same frequency-hopping patterns and spreading code (Pöpper et al. 2009).

Research work has been done to overcome the concern of establishing jamming-resilient communication with a pre-shared secret key. Strasser et al. proposed an anti-jamming system based on an uncoordinated frequency hopping (UFH), which enables two wireless devices to establish a secret key during the transmission time in the presence of a jammer (Strasser et al. 2008). Pöpper et al. created a scheme called uncoordinated direct-sequence spread spectrum (UDSSS), which can remove the requirement of a shared secret key for DSSS systems (Pöpper et al. 2009), and a sender always selects random spreading code sequences from a public set to spread a message.

Randomized Differential DSSS (RD-DSSS) is another anti-jamming system developed to remove the requirement of a shared secret key. In RD-DSSS, both a sender and a receiver share public spreading code sequences. A sender spreads a message based on a chosen index code, which is appended to the end of a spread message to facilitate the decoding at the receiver (Liu et al. 2010b). Liu et al. developed a new scheme, named Delayed Seed-Disclosure DSSS (DSD-DSSS), against jamming attacks with no need of sharing a secret key. They achieve anti-jamming wireless communication by generating a random seed, which can be used to generate random code sequences to spread each message. To disclose a seed, a sender spreads a seed using publicly known code sequence set and positions it to the end of a spreading message (Liu et al. 2010a).

Background Information on FHSS

FHSS transmits the wireless signal over multiple channels to avoid interference. Specifically, the transmitter and the receiver hop from one frequency to another based on a shared frequency-hopping pattern generated by a

pseudo-noise (PN) generator. The transmitter multiplies the transmit signal with a carrier of different frequencies, and the receiver searches signals according to the corresponding frequencies.

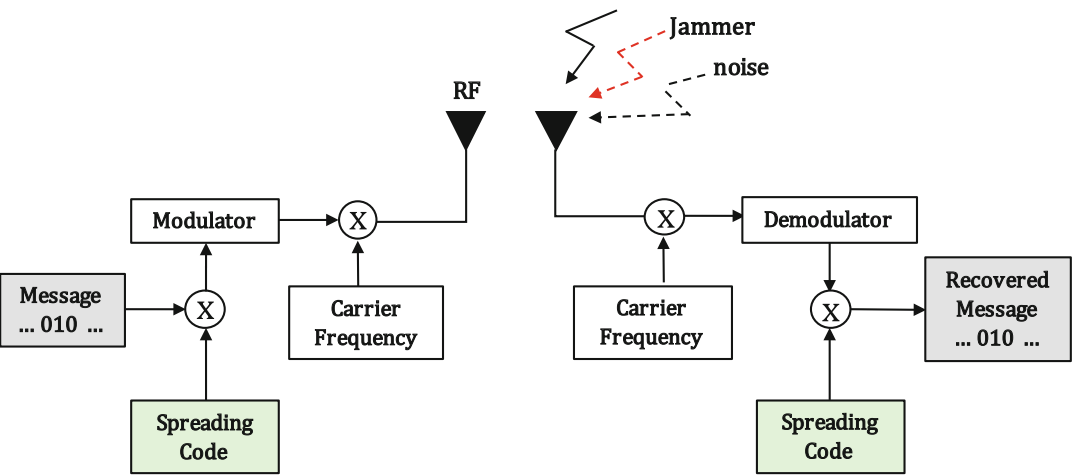
A receiver must be able to align with the signal sent by the transmitter. Toward this goal, both the transmitter and the receiver may stay on a certain channel without hopping until they hear from each other, or the transmitter may stay on a certain channel, and the receiver searches over different channels until it discovers the signal from the transmitter. Once the signal alignment is done, the transmitter and the receiver can start hopping over different frequencies according to the shared frequency-hopping pattern. Strasser et al. proposed a new way for signal alignment by using uncoordinated hopping and public key cryptography (Strasser et al. 2008). The basic idea is to let the transmitter and receiver randomly hop over different frequencies but at different hopping rates. Once the communicators are on the same channel, the transmitter sends the receiver a secret value that is encrypted by the public key of the receiver, who will decrypt this value using the corresponding private key. The transmitter and the receiver then generate shared hopping pattern based on the secret value. The advantage of this method is that the transmitter and the receiver are no longer required to have a

shared secret key prior to the wireless communication.

Background Information on DSSS

The concept of DSSS is more complicated than FHSS. DSSS is a modulation technique that can be applied to a baseband signal to increase its bandwidth. This can be achieved by multiplying an original signal by a spreading code formed by chips. This spreading code is considered as a pre-shared secret key between a sender and a receiver to expand a signal. Each spreading code is represented by a sequence of bits with values of 1 and -1 or 1 and 0 as polar or nonpolar representations, respectively (Goldsmith 2005; Simon et al. 1994). Each bit of an original message is multiplied with a spreading code, and the result is the spreading message to be transmitted through the wireless channel. A receiver does the reverse steps of a sender to despread a message. Figure 1 shows a basic DSSS communication system with two functions, including spreading and despreading.

An example of spreading and despreading a message For instance, assume that “10” is the original message and $+1 - 1 - 1 + 1$ is the spreading code used to spread the original message.



Wireless Jamming Attack, Fig. 1 Basic DSSS communication system

The sender first converts the original message “10” from a nonpolar to polar form, i.e., “10” is converted to $+1 - 1$. Then, the sender multiplies each bit of the original message with a spreading code to generate the spreading result. For this example, the result is $+1 - 1 - 1 + 1 - 1 + 1 + 1 - 1$.

To despread a received message, the receiver needs to correlate the received message with a local replica of the spreading code. In this example, the received message is $+111+11+1+11$, and the local replica of the spreading code is $+1 - 1 - 1 + 1$. The receiver aligns $+1 - 1 - 1 + 1$ with the first 4 chips of the received message (i.e., $+111 + 1$) and correlates them to get bit $+1$ (i.e., “1” in nonpolar form) and then aligns $+1 - 1 - 1 + 1$ with the next 4 chips of the received message (i.e., $-1 + 1 + 1 - 1$) and correlates them to get bit -1 (i.e., “0” in nonpolar form).

Synchronization In DSSS, identifying the beginning of a received message from the incoming bit stream is important for a receiver to recover the original message. This can be achieved by taking advantage of the autocorrelation property of the spreading code. A receiver computes the correlation between the received message and the spreading code in a sliding window way, and high correlation indicates the beginning of a message (Simon et al. 1994; Stallings 2006).

Key Applications

Jamming techniques are normally applied in tactical scenarios to block the wireless communication among the opponents. They can also be used in civilian applications to disable cell phone devices for certain purposes, e.g., preventing cheating in exams.

Cross-References

- [Wireless Data Security](#)
- [Wireless Key Establishment](#)

References

- Goldsmith A (2005) *Wireless communications*. Cambridge University Press, New York
- Grover K, Lim A, Yang Q (2014) Jamming and anti-jamming techniques in wireless networks: a survey. *Int J Ad Hoc Ubiquitous Comput* 17(4):197–215
- Liu A, Ning P, Dai H, Liu Y, Wang C (2010a) Defending DSSS-based broadcast communication against insider jammers via delayed seed-disclosure. In: *Proceedings of the 26th annual computer security applications conference*. ACM, pp 367–376
- Liu Y, Ning P, Dai H, Liu A (2010b) Randomized differential DSSS: jamming-resistant wireless broadcast communication. In: *INFOCOM, 2010 proceedings IEEE*. IEEE, pp 1–9
- Pöpper C, Strasser M, Capkun S (2009) Jamming-resistant broadcast communication without shared keys. In: *USENIX security symposium*, pp 231–248
- Simon MK, Omura JK, Scholtz RA, Levitt BK (1994) *Spread spectrum communications handbook*, revised edition. McGraw-Hill, Inc., New York
- Stallings W (2006) *Data and computer communications*, 8th edn. Prentice-Hall, Inc., Upper Saddle River
- Strasser M, Pöpper C, Capkun S, Cagalj M (2008) Jamming-resistant key establishment using uncoordinated frequency hopping. In: *Security and privacy, IEEE symposium on SP 2008*. IEEE, pp 64–78

Wireless Key Establishment

Tao Wang¹ and Yao Liu²

¹Department of Computer Science, New Mexico State University, Las Cruces, NM, USA

²Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

Synonyms

[Channel reciprocity](#); [Privacy amplification](#); [Quantization](#); [Reconciliation](#); [Spatial uncorrelation](#)

Definitions

Wireless key establishment is a method that allows two legitimate devices to establish a shared key utilizing the randomness inherent

in the wireless channel. In wireless context, the receiver and the transmitter of one wireless link will observe the same channel simultaneously. Such reciprocity property enables two legitimate devices to extract the same characteristics of the channel and agree on the same shared key. Meanwhile, due to the spatial uncorrelation of wireless channels, for two transmitters at different locations, the channels observed by the same receiver are different. Thus, the key established between a pair of wireless devices is confidential to a third un-located party.

Historical Background

Wireless devices have been widely used due to its remarkable evolution in the past two decades. Unlike traditional communication, a wireless device can communicate with any other device within its power range. This makes wireless communication vulnerable to potential attacks, because any devices within the power range of a wireless transmitter can receive the signal from this transmitter through the open public air. Therefore, eavesdropping becomes one of the major security problems to wireless systems. Intuitively, the communicators would like to share a common secret key, so that their conversation can be encrypted against eavesdroppers.

Traditional approaches to generate shared secret keys are mainly based on key pre-distribution schemes and public key cryptography. In the former, a centralized trusted third party generates, maintains, and distributes shared keys to communicators. However, such a scheme introduces a high key management complexity for large-scale wireless networks, which usually involve a large size of key pool and require intensive key distribution to support the key establishment between every pair of nodes.

On the other hand, public key cryptographic approaches to generate secret keys normally require the communication entities to be equipped with desired computing chips or modules, and they have been long regarded as

expensive in terms of computational complexity. For example, Diffie-Hellman algorithms, the most popular cryptographic tools to establish a shared secret key between two parties, require exponential operations on large numbers. Thus, they are not suitable for establishing the secret key among low-end wireless devices (e.g., wireless sensor nodes) that are of limited battery lifetime and computational capability.

Alternative key establishment techniques that are efficient and unique for wireless systems have been explored. There is an increasing concern about achieving fast and efficient key establishment via exploiting physical layer characteristics of wireless channels. The basic idea is to use the wireless channel reciprocity, which means that the receiver and the transmitter of one wireless link observe the same channel simultaneously. The reciprocity property of wireless channels allows two legitimate devices to establish a shared key. Due to the spatial uncorrelation of wireless channels, for two transmitters at different locations, the channels observed by the same receiver are different. Thus, the key established between a pair of wireless devices is confidential to a third un-located party. Past measurement results show that a distance of half a wavelength between the third party and the communicators can cause a mismatch about 50% between the channels observed by the third party and legitimate communication parties (Mathur et al. 2011).

Wireless channel reciprocity-based key establishment has formed a fruitful area, and it overcomes the drawbacks of traditional key establishment approaches due to its reduced computational effort and relaxed key management requirement.

Foundations

In general, a shared key is a binary bit sequence that is only known to the transmitter and the receiver. They use the shared key to encrypt and decrypt transmitted information to secure their communication. In channel reciprocity-based key establishment, a shared key is generated

from one or more channel characteristics, such as signal frequency-phase, received signal strength (RSS), and channel impulse response (CIR). It normally takes three steps to implement a channel reciprocity-based key establishment, and they are quantization, reconciliation, and privacy amplification (Wang et al. 2015). In quantization, the transmitter and receiver first sample the transmitted signal at a certain frequency, and then both of them quantize the sampled signal based on particular thresholds to generate initial binary bit sequences. Due to imperfect reciprocity and random noise, the bit sequence generated at the transmitter and the receiver from quantization may not be exactly the same. Hence, reconciliation schemes will be applied to deal with these mismatch bits. The objective of reconciliation is to correct or delete these mismatch bits using minimum channel information. After reconciliation, the transmitter and the receiver then perform privacy amplification to avoid a malicious adversary deducing the secret bit sequence. Existing approaches on wireless channel reciprocity-based key establishment mostly focus on the following critical aspects:

1. **Quantization:** Quantization is the most important part of the shared key establishment, because it provides initial information of the wireless channel. All the remaining steps expect an efficient and precise quantization output. The essential challenge of quantization is how to determine channel metrics that can fully characterize a unique wireless channel. The selection of channel metrics has a direct impact on the performance of quantization. Appropriate channel metrics (e.g., RSS, CIR, and frequency-phase information) can be used to describe a particular wireless channel and achieve a desired performance (Liu et al. 2012; Mathur et al. 2008; Ali et al. 2010; Zhu et al. 2013). Note that a good quantization performance also depends on the choice of quantization thresholds. A single threshold will lead a low generation rate, while multilevel thresholds are susceptible to the random noise.
2. **Reconciliation and privacy amplification:** Information reconciliation and privacy amplification are two indispensable steps of the wireless key establishment. Both steps collaborate with each other to achieve an adequate key generation rate with minimum information leakage. In reconciliation, devices need to exchange some channel information to correct mismatch bits. Thus, the goal of reconciliation is to minimize the channel information exchanged between the communicators, so that the chance of information leakage can be reduced. Similarly, privacy amplification aims at “amplifying” the difficulty of an eavesdropper to deduce the shared secret based on the channel information exchanged in reconciliation.
3. **Feasibility and security:** It is essential to analyze the feasibility of wireless channel reciprocity-based key establishment when they are used in a practical scenario. Investigations on the requirement they should satisfy to guarantee the generation of secret keys have been conducted since 1993 (e.g., Maurer 1993; Ahlswede and Csiszár 1993). On the other hand, using wireless channel reciprocity information to establish a shared secret key is still a developing field, and multiple attacks against such techniques have been discovered. For example, an eavesdropper may launch proximity attack (e.g., Mathur et al. 2011), in which the eavesdropper infers the secret key established between the target communicators through physically approaching to the communicators or predicting the channel characteristics between them.

Key Applications

Wireless key establishment can be applied in wireless networks especially with low-end devices (e.g., wireless sensor networks, Internet of Things) to facilitate key negotiations and preserve the confidentiality of communications between two legitimate devices.

Cross-References

- ▶ [Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks](#)
- ▶ [Authentication](#)
- ▶ [Group Key Agreement for Wireless Networks](#)
- ▶ [IoT Security](#)
- ▶ [Mobile Security and Privacy](#)
- ▶ [Wireless Group Authentication](#)

References

- Ahlswede R, Csiszár I (1993) Common randomness in information theory and cryptography. I. Secret sharing. *IEEE Trans Inf Theory* 39(4):1121–1132
- Ali ST, Sivaraman V, Ostry D (2010) Secret key generation rate vs. reconciliation cost using wireless channel characteristics in body area networks. In: 2010 IEEE/IFIP 8th international conference on embedded and ubiquitous computing (EUC). IEEE, pp 644–650
- Liu Y, Draper SC, Sayeed AM (2012) Exploiting channel diversity in secret key generation from multipath fading randomness. *IEEE Trans Inf Forensics Secur* 7(5):1484–1497
- Mathur S, Trappe W, Mandayam N, Ye C, Reznik A (2008) Radio-telepathy: extracting a secret key from an unauthenticated wireless channel. In: Proceedings of the 14th ACM international conference on Mobile computing and networking. ACM, pp 128–139
- Mathur S, Miller R, Varshavsky A, Trappe W, Mandayam N (2011) Proximate: proximity-based secure pairing using ambient wireless signals. In: Proceedings of the 9th international conference on Mobile systems, applications, and services. ACM, pp 211–224
- Maurer UM (1993) Secret key agreement by public discussion from common information. *IEEE Trans Inf Theory* 39(3):733–742
- Wang T, Liu Y, Vasilakos AV (2015) Survey on channel reciprocity based key establishment techniques for wireless systems. *Wirel Netw* 21(6):1835–1846
- Zhu X, Xu F, Novak E, Tan CC, Li Q, Chen G (2013) Extracting secret key from wireless link dynamics in vehicular environments. In: INFOCOM, 2013 Proceedings IEEE. IEEE, pp 2283–2291

Wireless Link Scheduling

- ▶ [Interference-Aware Scheduling in Wireless Networks](#)

Wireless Network

- ▶ [Power Control in Wireless Networks](#)
- ▶ [Random Distance Distribution](#)
- ▶ [Trustworthiness Evaluation](#)
- ▶ [Ubiquitous Big Data](#)
- ▶ [Wireless Jamming Attack](#)

Wireless Recharging Technology

- ▶ [Energy Harvesting Technologies in Wireless Sensor Networks](#)

Wireless Relay Channel

- ▶ [Fundamental Properties of Wireless Relays and Their Channel Estimation](#)

Wireless Relay System

- ▶ [Fundamental Properties of Wireless Relays and Their Channel Estimation](#)

Wireless Resource Virtualization

- ▶ [Resource Scheduling in Virtualized Wireless Networks](#)

Wireless Sensor Network

- ▶ [Quality of Service of Wireless Sensor Networks Modeled by Ginibre Point Processes](#)

Wireless Sensor Network Security

Bobby Sharma
School of Technology, Assam Don Bosco
University, Guwahati, India

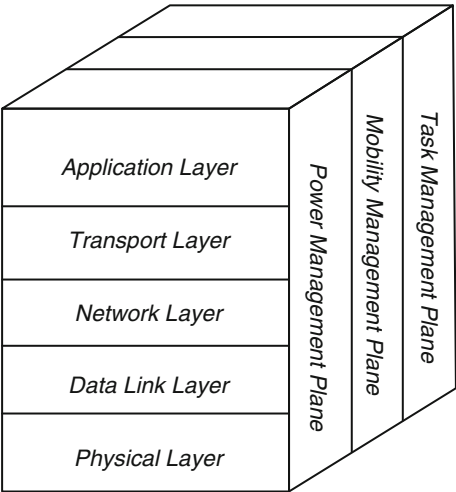
Synonyms

Consistency, integrity; Definition of wireless sensor network; Insider, internal; Outsider, external; Threat, attack; Transmitting, passing forwarding

Wireless Sensor Network and Security Issues

Wireless sensor network (WSN) is a kind of network in which several numbers of tiny nodes are deployed in such a way that they can monitor, collect data from environment, process them, and pass them using radio signal (Carlo 2014; Marco 2011; Bhaskar 2005). These are self-configured. In such situation, collaborative effort of all nodes in WSN is highly expected. Mostly sensor nodes are deployed in hard to reach area where it is very difficult to have human intervention. These are auto operated; hence, special care must be taken for energy consumption as well as external attack Scalabili WSN is high (Jaydip 2012). It provides bridge between real world and virtual world (Marco 2011). Key Application of WSN by looking the versatility and flexibility of WSN, these are implemented in various fields of applications including agriculture, medical, engineering, environmental science, water resource management, industrial engineering, security and surveillance system, transportation, smart grid application, etc. (Bhaskar 2005); in spite of having so many scopes for implementation, these are suffering from some limitations including proper security features.

Figure 1 shows the different layers of WSN. In most of WSN, basically five different layers are available, namely, application layer, transport layer, network layer, data link layer, and physical



Wireless Sensor Network Security, Fig. 1 Layer structure of WSN (Kavitha and Sridharan 2010)

layer. Functions of these layers are more or less same as that of any other network layers. Apart from these, another three cross layers are also available to take care of special function of WSN. These are, namely, power management layer, mobility management layer, and task management layers.

WSN security primarily includes prediction, detection, and prevention of unauthorized access of network, stealing, modification, or disruption of resources, disturbing users from unnecessary mail, pop-up, monitoring user activities without authorization, and more consumption of network resources such as bandwidth, power backup, etc.

It also includes protection of different threats and attacks. Even while protecting the threat and attack, it must be concerned that it should not disrupt the normal activities of the network. Otherwise due to lack of inefficient security services, the user may face lots of problem.

People are working on different ways to prevent WSN from all sorts of security faults. Some of the ways are given below.

Passive network security devices understand the malicious network traffic, and accordingly it prevents the traffic and save the network from attacks.

But the active security devices resist the specific network traffic to keep the network function intake.

With the help of preventive security devices, network administrator can understand the potential threat and alert the user to take preventive measures.

Applications or devices are built in such a way that they understand potential attack, unauthorized network traffic, or any other threat that network may face.

Specifically for sensor network, there must be auto protected system as it is quite difficult to maintain the same manually on regular basis. While transmitting data from various sensor nodes to base station, it should be in encrypted format so that external attacker is not able to get data easily. Any partially or fully disabled sensor nodes should be remotely blocked.

Attacks in WSN

Types of attack that are found in the case of WSN are shown below (Vikash et al. 2014).

Due to unique characteristics of WSN and lack of constraints in hardware and software, WSN is very much vulnerable to different kinds of attack (Pathan et al. 2006). Sensor nodes communicate among themselves over wireless link; it leads them to eavesdrop on. An attacker may easily push the malicious data into the network. Denial-of-service attack may take place easily in WSN (Pathan et al. 2006; Anwar et al. 2014; Kavitha and Sridharan 2010; Vikash et al. 2014).

Attacks are also defined as outsider as well as insider attack. In case of outsider attack, attackers are not a part of WSN. It may cause passive eavesdropping on data transmission, or it may inject malicious data into the network with a motive to consume network resources.

But in case of insider attack, legitimate nodes of WSN may compromise with attackers, or independently it may attack the system. These are more dangerous because the trusted nodes of the network may break the authentication system and exposed the confidentiality of the network. It is very difficult to understand the attack pat-

tern. Even the attackers suppress the regular data flow from various nodes to BS. Due to packet-dropping attack of this kind, network performance is gradually coming down (Kavitha and Sridharan 2010; Kahina 2015).

WSN is suffering from both active and passive attack. Active attacks are kinds of attack in which attackers are silently monitoring the system, data forwarding, channel blocking by flooding packets in the network, diverting data, dropping packets, routing disorder, malfunctioning the nodes, etc. Here attackers take the control over the network. Passive attacks are kinds of attack in which attackers are listening or monitoring the traffic, analyzing traffic, stealing authentication-related data, breaking data encryption system, disclosing data, etc. (Vikash et al. 2014; Kahina 2015).

Basic classifications of attack are (Vikash et al. 2014):

1. Active Attack
 - 1.1 Routing Attacks in Sensor Networks
 - 1.1.1 Attacks on Information in Transit
 - 1.1.2 Selective Forwarding
 - 1.1.3 Black Hole/Sinkhole Attack
 - 1.1.4 Wormhole Attacks
 - 1.1.5 Hello Flood Attack
 - 1.2 Denial of Services
 - 1.3 Node Subversion
 - 1.4 Node Malfunction
 - 1.5 Node Outage
 - 1.6 Physical Attacks
 - 1.7 Message Corruption
 - 1.8 False Node
 - 1.9 Node Replication Attacks
 - 1.10 Passive Information Gathering
2. Passive Attack
 - 2.1 Monitor and Eavesdropping
 - 2.2 Traffic Analysis
 - 2.3 Camouflage Adversaries

Routing attack is a network layer attack. Different kinds of attack under this category include. In case of attacks on information in transit, the attacker monitors the traffic, and as per requirement, they can spoof, alter, and replay the traffic to modify, interrupt, or intercept the packet.

Though it is expected that sensor nodes should forward the traffic, yet in some cases, nodes may compromise with attacker and selectively forward the packets instead of forwarding all data.

In case of blackhole attack, attacker drags all the traffic toward it. As a result of this destination is failure to get data.

In wormhole attack, attacker replays the broadcast request packets that it receives from source to its neighbors and convinces them in such a way that as if the source is one hop away from it, thus tunneling the packets in different directions.

In Hello flood attack, the attacker sends high-powered Hello messages to its neighbors. By this, adversaries advertise themselves to the neighbors so that neighbors understand that path through the attacker is superior than any other available paths. This attack is a very dangerous attack because it consumes high amount of network resources which ultimately leads to low network performance.

Denial-of-service (DoS) attack may be data link layer attack or network layer attack. Here, attacker may inject fake bogus packets in the channel to force security measures. As a result of this, network fails to maintain the network primitives like availability, integrity, and confidentiality.

Sometimes, some sensor nodes of the network may compromise with the adversaries and disclosing the security key to the attacker. Hence confidentiality of the network is not maintained. Such kind of attack is categorized under node subversion.

Node malfunctioning may corrupt the entire system by generating unnecessary data to lose the data integrity of the system.

In some situation, sensor node may suddenly fail to work due to physical, location, or software problem. But in case of outage attack, the attacker intentionally creates the situation to break down the ongoing data path.

In physical attack, attacker may destroy any hardware of the sensor node by tempering the circuitry.

In message corruption attack, original message is tempered with malicious data.

In false node creation attack, attacker creates false node in the network, misleading the genuine nodes. Thus data delivery process as well as network resource is getting down.

In node replication attack, attacker replicates the existing node by identifying sensor node's ID. This attack may quickly degrade the network resources like bandwidth, battery power of the nodes, etc. which are considered as major concern for sensor network.

Some malicious nodes may silently monitor the network traffic and collect required data from the network. This is categorized under passive information gathering attack.

In case of passive attack, also adversaries monitor the traffic and snooping data during communication. In traffic analysis attack, attacker analyzes and tempers data accordingly. Camouflage adversaries may focus themselves as an actual node in the network and divert the packets to different directions.

Basic Security Scheme for WSN

Unlike any other network, any kind of security schemes cannot be implemented in WSN. The tiny nodes in WSN are designed in such a way that they consume less energy. Irrespective of their configuration, if any kind of security features are implemented then it may lose the basic network primitives such as authentication, integrity, privacy, nonrepudiation, and anti-playback. Some of the security features that are implemented in WSN are as follows:

(a) Cryptograph

Data encryption and decryption technology for a wired or other wireless network are not the same as that of implementation in WSN (Pathan et al. 2006). WSN contains tiny nodes which are very much concern about resources. At the same time, cryptographic algorithm is resource consuming. Since WSN suffers from several limitations including battery power, small memory, etc., traditional cryptographic algorithm is quite difficult to implement (Pathan et al. 2006; Rawya

and Yasmin 2015; Madhumita 2014). Of course, various cryptographic algorithms are proposed to meet the security requirements including authenticity, confidentiality, and data integrity (Rawya and Yasmin 2015). There are two different ways to protect the system, namely, symmetric and asymmetric algorithms. Some examples are AES (Advanced Encryption Standard), ECC (Elliptic Curve Cryptography), RC4 (Rivest Cipher 4), RC5 (Rivest Cipher 5), RSA (Ron Shamir and Adleman), etc. (Daniel et al. 2017). Symmetric algorithm is cost-effective and secure enough. Of course, sharing of secret key may be an issue. Though asymmetric algorithm is able to resolve the problem of sharing of secret key, these are more resource consuming and slow. In some cases, hybrid algorithm is proposed (Rawya and Yasmin 2015). In Rawya and Yasmin (2015), authors designed an algorithm in which two phases are included; in phase I, both symmetric and asymmetric algorithms like AES and ECC are used. In phase II, XOR-DUAL RSA is used for security in addition to hashing and MD5 for data integrity.

(b) Steganography

In contrast to cryptography, steganography hides the message existence (Pathan et al. 2006; David and Hervé 2012). This is one of the best methods to secure data as external users are generally unaware about the existence of the message (Reetika and Anshul 2017). In this method, carrier is modified so that it is not perceptible. In Reetika and Anshul (2017), authors propose a system in which steganography technique is used for data hiding in WSN, and then various parameters like MSE, PSNR, AMD, energy consumption, network lifetime, etc. are calculated.

(c) Physical Layer Secure Access

It is a method which uses frequency hopping. The method uses available frequency set, dwell time, and hopping pattern. Efficient design is one of the most important parts

of this method. Also the synchronization of sender and receiver is another important issue (Pathan et al. 2006). It is for intrinsic security process induced based on physical layer capabilities like SNR (signal-to-noise ratio) gap, intended jamming, randomness of channel, etc. (Jinho et al. 2013). This security feature tries to hide the existence of certain nodes though ongoing communication is available. This process works fine unless and until the existence of hidden nodes is not disclosed to adversaries. Some of the existing physical layer security approaches included secrecy capacity, channel signature, spectrum spreading of signal energy, cooperation, etc. (Tommaso et al. 2016).

(d) Watermark

Watermarking technique has been widely used in intellectual property protection like text, audio/video, circuit design, and data security in sensor network (Jennifer et al. 2016). Due to lack of some limitations of cryptographic security solutions, it is identified that watermarking security solutions are proposed due to its light resource requirements (Bambang et al. 2012).

Security Challenges

Security challenges are very important in WSN. Some of the security challenges available for WSN are given below (Daniel et al. 2017; Madhumita 2014; Raja et al. 2014).

Authenticity: Any intruder may be part of the network. Like any other nodes available in WSN, it may also try to penetrate malicious data into the network. Hence, anything coming from authorized nodes must be verified so that the network can maintain its security.

Availability: While implementing security features, the user must concern about the availability of sensor nodes. There may be two different points of concern: firstly, the network must understand about the availability of all nodes. If not, it must find the valid cause. Secondly, after

implementation of security features, nodes must work like normal nodes.

Confidentiality: It is another important point to be concerned so that the data should not be disclosed to any intruder.

Freshness: Security measures also consider the point of data freshness. On time data delivery and availability of data is one of the major points of concern.

Integrity: Data integrity is a point where data consistency as well as data privacy is maintained so that attacker cannot modify, delete, or drop the data.

Localization: Localization implies the sending and receiving of expected data to and from the actual nodes.

Conclusion

From various studies, it has been seen that WSN has vast applications in current scenarios. It has some unique features which makes them different from other networks. Smart nodes in WSN are mostly collecting data from the environment, then process them, and send these data to BS. Due to its flexibilities in applications as well as easy implementation, these are suffering from various kinds of limitations including different attacks. Lots of research works are going on to detect and eliminate those attacks from WSN.

References

- Anwar R, Bakhtiari M, Zainal A, Abdullah A and Qureshi K (2014) Security Issues and Attacks in Wireless Sensor Network. *World Applied Sciences Journal* 30(10): SSN 1818-4952. 1224–1227. <https://doi.org/10.5829/idosi.wasj.2014.30.10.334>
- Bambang H, Vidyasagar P, Jaipal S (2012) Watermarking technique for wireless sensor networks: a state of the art. In: 2012 Eighth international conference on semantics, knowledge and grids. <https://doi.org/10.1109/SKG.2012.56>
- Bhaskar K (2005) An introduction to wireless sensor networks. Tutorial presented at the second international conference on intelligent sensing and information processing (ICISIP)
- Carlo F (2014) An introduction to wireless sensor networks, Draft, version 1.8
- Daniel C, Solenir F, Gledson O (2017) Cryptography in wireless multimedia sensor networks: a survey and research directions. *Cryptography* 1:4. <https://doi.org/10.3390/cryptography1010004>
- David M, Hervé G (2012) Steganography in MAC layers of 802.15.4 protocol for securing wireless sensor networks. HAL Id: hal-00661843, IWNS 2010, 2nd IEEE Int. Workshop on Network Steganography, 2010, \break{} China
- Jaydip S (2012) Security in wireless sensor networks. *Wireless sensor networks: Current Status and Future Trends*, Shafiullah Khan, Al-Sakib Khan Pathan, and Nabil A. Alrajeh (Eds.), Auerbach Publications, CRC Press, Taylor & Francis Group, USA
- Jennifer W, Jessica F, Darko K, Miodrag P (2016) Security in sensor networks: watermarking techniques. *Wireless sensor networks*, pp. 305–323. https://doi.org/10.1007/978-1-4020-7884-2_14
- Jinho C, Jeongseok H, Hyoungseok J (2013) Physical layer security for wireless sensor networks. In: 2013 IEEE 24th international symposium on personal, indoor and mobile radio communications: fundamentals and PHY track
- Kahina C (2015) Security issues in wireless sensor networks: attacks and countermeasures. In: *Proceedings of the world congress on engineering 2015*, vol I. WCE, 1–3 July 2015
- Kavitha T, Sridharan D (2010) Security vulnerabilities in wireless sensor networks: a survey. *J Inf Assur Secur* 5(2010):031–044
- Madhumita P (2014) Security in wireless sensor networks using cryptographic techniques. *Am J Eng Res* 03(01):50–56. e-ISSN 2320-0847, p-ISSN 2320-0936
- Marco Z (2011) Introduction to wireless sensor network. Trieste-Italy page 2
- Pathan A, Woo Lee H, Hong S (2006) Security in Wireless Sensor Networks: Issues an Challenges. supported by MIC and ITRC Project. ICACT2006, SBN 89-5519-129-4, 20–22
- Raja A, Majid B, Anazida Z, Abdul A, Kashif Q (2014) Security issues and attacks in wireless sensor network. *World Appl Sci J* 30(10):1224–1227. <https://doi.org/10.5829/idosi.wasj.2014.30.10.334>. ISSN 1818-4952
- Rawya R, Yasmin A (2015) Two-phase hybrid cryptography algorithm for wireless sensor networks. *J Electr Syst Inf Technol* 2(2015):296–313
- Reetika S, Anshul S (2017) Statistical steganography technique for minimizing MSE in wireless sensor networks. *Int J Eng Pure Appl Sci* 2(2). ISSN NO. 2456-3129
- Tommaso P, Luca B, Lorenzo M (2016) The role of physical layer security in IoT: a novel perspective. *Information* 7:49. <https://doi.org/10.3390/info7030049>
- Vikash K, Anshu J, Barwal P (2014) Wireless sensor networks: security issues, challenges and solutions. *Int J Inf Comput Technol* 4:859–868. ISSN 0974-2239

Wireless Sensor Networks

- [Application of Machine Learning in Wireless Sensor Network](#)
- [Energy Harvesting Sensor Node Scheduling in Wireless Sensor Networks](#)
- [Wireless Sensor Networks Enabled Urban Traffic Management](#)

Wireless Sensor Networks Enabled Urban Traffic Management

Zhaolong Ning and Xiaojie Wang
Dalian University of Technology, Dalian, China

Synonyms

[Traffic management](#); [Vehicular networks](#); [Wireless sensor networks](#)

Definition

An effective urban traffic management system intends to manage traffic effectively by leveraging wireless sensor networks enabled processing technologies and intelligent system algorithms for smart cities.

Historical Background

Generally speaking, Wireless Sensor Networks (WSNs) consisting of a group of resource-constrained sensors to collect data from surrounding environments are becoming key components in the emerging areas of cyber physical systems. On one hand, energy efficiency is crucial for most sensors due to battery supply. On the other hand, topology management and network coverage are also significant in

WSNs (Haque and Abu-Ghazaleh 2016). In general, the main research background of WSN-enabled traffic management is involving from capacity increase, mobility support to quality of experience improvement.

The study of Internet of Vehicles (IoVs) has sprung up, where vehicles are viewed as sensor hubs to capture information by in-vehicle sensors for kinds of mobile applications. Based on the research report from BI Intelligence in 2015, 75% of cars will be equipped with network devices to enable Internet access on roads by 2020. The deep integration of intelligent sensors and communication technologies opens up a new frontier for IoVs in smart cities. Thanks to the development of communication and information technologies, connected vehicles in IoVs are playing a crucial role for urban traffic management, such as road safety improvement and transportation efficiency enhancement.

The rapid development of WSNs in vehicular networks enables high-efficient traffic management. A comprehensive survey of traffic management system in smart cities is conducted in Djahel et al. (2015), from the aspects of information collection to that of service delivery. The data sensing and gathering phase can measure traffic parameters, e.g., traffic volume and speed, through road-monitoring equipment. After that, the generated information is periodically reported to a management entity. Because the data are fused and aggregated during the data fusion, processing, and aggregation processes, how to extract useful information for traffic management is significant. The data exploitation phase employs the obtained knowledge through the data processing phase to select optimal routes for vehicles, forecast short-term traffic, and provide road-traffic statistics. In the service delivery process, the traffic management system forwards the acquired knowledge to terminal users through devices, e.g., smartphones and vehicle onboard units.

Due to the high mobility of vehicles equipped with sensors, traffic management in urban areas still faces several challenges. Sink mobility management is generally regarded as an effective method to promote performances in WSNs, such

as it can relieve the traffic burden for a group of network nodes. The corresponding sink mobility management schemes can be classified into four categories: uncontrollable mobility, path-restricted mobility, location-restricted mobility, and unrestricted mobility (Gu et al. 2016). The behavior of mobile sinks has direct impact on the performance of uncontrollable mobility. Its main challenge is lacking of sink information to update sink location and forward information. Path-restricted mobility is related to the geographic constraints, i.e., vehicles attached by sinks collect data by moving along certain roads. Under this circumstance, how to collect data, such as speed and routing information is challenging. Location-restricted mobility has a direct relationship with the sink location, i.e., sinks merely stop at some locations for data collection. Different from path-restricted mobility where the sink trajectory is predetermined, and location-restricted mobility where the stops are fixed, unrestricted mobility generally embraces the issues of not only data routing but also motion control.

For the sake of realizing high-efficient urban traffic management, three critical issues deserve to be investigated: (1) Real-time traffic condition perception: The various high-resolution roadside and on-board sensors require to be deployed to sense real-time traffic circumstances, such as location, speed, road throughput, and weather condition; (2) Low-latency communication and massive-data storage: Sensors equipped by vehicles generate vast amount of data with distinct data types, sizes, and dimensionality, challenging the communication bandwidth, data processing speed, and storage capability of current vehicular networks; (3) Traffic prediction and real-time responsiveness: The generated traffic data from various sources can be leveraged to monitor traffic density and road throughput in a real-time manner, so that prompt decisions can be made according to the network prediction to guide traffic flows. Some open issues are how to design a context-aware system for traffic light control to reduce the waiting time at road intersections, how to construct a real-time response system for accident rescue, and how to design an accurate model to estimate network

traffic according to the obtained massive traffic data (Nellore and Hancke 2016).

Key Applications

WSN-enabled urban traffic management is promising for smart cities, and some applications have been well discussed by the previous researches. In this section, we focus on some emerging applications for urban traffic management.

Deep Learning-Based Traffic Management

In recent years, deep learning has become a breakthrough methodology in machine learning for distinct research fields in computer science and other disciplines. However, it should be noticed that the recent endeavors to train deep learning architectures for network traffic control generally concentrate on some benchmark routing protocols, which cannot cope with the increasingly complex network circumstances due to the lack of computing intelligence (Fadlullah et al. 2017). To overcome these limitations, the integration of deep learning and network traffic optimization is promising to make intelligent routing decision for message forwarding. Furthermore, effective traffic control in wireless backbone networks is also significant. The reasons are the traditional routing schemes with network abnormality cannot learn their knowledge from previous experiences. Intelligent network traffic control enabled by real-time deep learning is advocated, which can integrate deep convolutional neural networks with unique input and output characters to describe the character of network backbone.

Another key application for deep learning-based traffic management is automated vehicles (acknowledged as driverless cars or self-driving vehicles as well), thanks to the rapid development of advanced sensors in situational awareness and autonomous navigation systems. However, various sensors equipped on vehicles need to be integrated, challenging the design of machine learning-based traffic management for real-time driving decisions. Recently, deep neural networks

have also been leveraged to analyze the generated multi-modal information by sensors for automated vehicles.

Smart Parking

Due to the ever-increasing volume of traffic congestion, smart parking has become a needpress from the viewpoint of both research fields and economic interests. By leveraging the advanced sensors to collect information for communication, drivers can find suitable parking spaces effectively. Generally speaking, smart parking solutions include information collection, system deployment, and service dissemination. A comprehensive overview of the available sensing technologies can be provided by the information collection to check the current empty parking space of parking lots. Among the WSN-enabled smart parking solutions, how to guarantee the identification and transmission of vehicular information are two main research topics. The identification of vehicular information depends on sensors for environment detection. The transmission of vehicular information relies on the integration of sensors and citywide infrastructures to collect information. Currently, crowdsensing and shared parking are two promising WSN-enabled urban services for smart parking. Software system exploitation and data analysis are the main concentration of system deployment. The corresponding software embraces guidance, reservation, E-parking, and information monitor for both managers and users. Driven by WSNs, the collected data can be leveraged to provide an analysis for urban traffic management. The relationship between information and social features reflects service dissemination. Due to individual behavior, information dissemination and parking competition are becoming two main issues. The former has a direct relationship with communication patterns among vehicles, while the latter happens due to the limitation of network resource.

Smart LED Streetlight

Owing to the development of low-power and low-cost WSNs, such as ZigBee, LoRa, and

Sigfox, our city is becoming smarter than ever before, where smart traffic, smart energy, and smart streetlight system become realized. A rather practical question is how to fuse the sensed data in a comprehensive manner for effective urban traffic management. It is well acknowledged that WSNs are very suitable for LED streetlight systems, which can be viewed as a core application for vehicles and pedestrians by providing low-cost and low power outdoor lighting. The combination of sensors and ZigBee-based wireless sensor modules is able to offer a platform for LED streetlight applications. In Daely et al. (2017), the authors designed a smart LED streetlight system by integrating public weather data, ZigBee-based wireless communication, and web-based management system for urban traffic management.

Energy Harvesting in Vehicular Networks

Energy harvesting is promising for energy-constrained wireless networks, such as WSNs and vehicular networks (Atallah et al. 2016). The former consists of sensors with limited lifetime, and the latter contains low-cost and battery-powered roadside units in rural areas or along highways. Through the conversion from energy to direct current voltage, energy harvesting becomes a research hotspot to prolong the battery lifetime of electronic devices. The corresponding applications enabled by WSNs cover traffic monitor, health monitor, and so on. Energy harvesting solutions can also be leveraged to operate and charge a large number of low-power devices. Radio frequency-based energy transfer and harvesting techniques are effective solutions to extend the lifetime of battery-based sensor nodes.

Smartphone-Based Vehicle Telematics

Smartphone is becoming a modern kind of sensors, which can offer various kinds of services and applications for our daily life. The integration of smartphones and vehicular networks is advocated for urban traffic management, which is not only cost effective but also can facilitate various components in urban traffic management

systems (Wahlström et al. 2017). However, the IEEE 802.11p interfaces are not equipped by commercial smartphones.

Although it is demonstrated that smartphones are not necessary to be compliant with the IEEE 802.11p standard, it is challenging to fulfill the sensing, computing, and storage capabilities of smartphones before forwarding the sensed data to an in-vehicle 802.11p device through Bluetooth or WiFi. The consumed energy by hardware during application running is a core factor for smartphones. Besides, black boxes installed by vehicles or vehicle's on-board system can acquire information, monitor, and feed back the performance of vehicles for traffic management. As the emerging of crowdsourcing/crowdsensing and sharing economy, how to fuse information from vehicles, pedestrians, and local infrastructure challenges the design of cooperative intelligent transportation systems, especially for individuals with limited sensing ability or within a limited geographical region.

References

- Atallah R et al (2016) Energy harvesting in vehicular networks: a contemporary survey. *IEEE Wirel Commun* 23(2):70–77
- Daely PT et al (2017) Design of smart LED streetlight system for smart city with web-based management system. *IEEE Sensors J* 17(18):6100–6110
- Djahel S et al (2015) A communications-oriented perspective on traffic management systems for smart cities: challenges and innovative approaches. *IEEE Commun Surv Tutorials* 17(1):125–151
- Fadlullah Z et al (2017) State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Commun Surv Tutorials* 19(4):2432–2455
- Haque IT, Abu-Ghazaleh N (2016) Wireless software defined networking: a survey and taxonomy. *IEEE Commun Surv Tutorials* 18(4):2713–2737
- Gu Y et al (2016) The evolution of sink mobility management in wireless sensor networks: a survey. *IEEE Commun Surv Tutorials* 18(1):507–524
- Nellore K, Hancke GP (2016) A survey on urban traffic management system using wireless sensor networks. *Sensors* 16(2):157
- Wahlström J et al (2017) Smartphone-based vehicle telematics: a ten-year anniversary. *IEEE Trans Intell Transp Syst* 18(10):2802–2825

Wireless Sensors Networks

- [Lattice Reduction and Its Applications in Wireless Sensors Network](#)

Wireless Technology in Data Centers

- [Wireless Communication in Data Center](#)

Wireless Video Delivery

Takuya Fujiihashi¹ and Takashi Watanabe²

¹Graduate School of Science and Engineering, Ehime University, Ehime, Japan

²Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

Synonyms

[Video delivery over wireless networks](#); [Wireless video transmission](#)

Definitions

Wireless video delivery refers to a video transmission technology over wireless channels and networks. A sender, e.g., content server, base station, and Wi-Fi access point, encodes video frames and transmits the encoded video frames over wireless links/networks. The transmitted video frames are impaired by losses and effective noise due to limited bandwidth and unstable channel quality. The receiver, e.g., laptop, smartphone, and tablet, receives the transmitted video frames and decodes the video frames for playback.

Historical Background

Video streaming over the Internet was born in 1993, and a commercial product of Internet video streaming was started in 1995. For realizing wireless multimedia delivery including video delivery, the first concept of the mobile multimedia device was shown by Sheng et al. (1992). One of the major issues in wireless video delivery is that video streaming requires high throughput to deliver high-quality video, while the capacity of wireless links/networks is considerably limited compared with wired links/networks. To discuss the feasibility of video streaming in band-limited wireless links/networks, video streaming over (very) low bandwidth links has been studied since 1994. For example, Khansari et al. (1994a) and Pelz (1994) consider the transmission of video frames in the Quarter Common Intermediate Format (QCIF) over the wireless links with an available data rate of 64 Kbits/s. They demonstrated the feasibility of wireless applications by using a video encoder/decoder with sophisticated signal processing and error correction techniques. From 1995, many video delivery techniques under the consideration of wireless link features, which are referred to as “cross-layer” techniques, have been studied for high-quality, reliable, and scalable wireless video delivery. Specifically, the existing cross-layer studies mainly tackle (1) efficient video delivery under limited throughput, (2) robustness to various channel conditions, and (3) scalability for multiple wireless devices.

Efficient Video Delivery Under Limited Throughput

As mentioned above, video delivery applications traditionally require a higher-throughput requirement due to a large amount of traffic. If the amount of video traffic is higher than the available throughput, it causes stalls during video playback and low user's satisfaction for wireless applications. To efficiently deliver video contents within an available throughput without stalls, existing studies on wireless video delivery mainly use two types of coding techniques to reduce the expected distortion of the received video contents

since 1994: scalable (layered) video coding and rate control.

For example, Khansari et al. (1994b) designed layered coding techniques for wireless video delivery. Specifically, they consider two paths (high and low rates) in end-to-end video session and send low- and high-frequency-domain coefficients in high-rate and low-rate path, respectively. Here, low-frequency coefficients provide baseline quality of video contents, while high-frequency coefficients enhance the quality of the video contents. Song and Chen (2007) proposed a joint application and physical layer design to transmit scalable video over multiple-input and multiple-output (MIMO) wireless systems. In this system, scalable video coding (SVC), such as H.264/SVC (Schwarz et al. 2007), generates layered bitstreams in application layer. At the physical layer, each subchannel's quality can be obtained by partial channel information. Finally, higher priority layer bitstream will be transmitted over higher-quality subchannel to realize unequal error protection for high-quality video delivery.

In the rate control, Ortega and Khansari (1995) proposed a rate control algorithm of the video encoder for variable bit rate in wireless links/networks. To realize better video quality under a variable bit rate, they use a channel model based on a Markov chain model for rate control to predict future channel behavior. From evaluations, they demonstrated the rate control algorithm with the estimated channel model keeps better video quality without actual channel information. Hsu et al. (1999) also proposed a rate control algorithm under the consideration of a time-varying burst loss. They consider the two-state Markov model and N-state Markov model for estimating burst loss into the rate control. Stankovic et al. (2004) designed two types of the bit allocation algorithms for forward error correction (FEC) systems in packet erasure and flat-fading channels. Specifically, the first algorithm provides Reed-Solomon (RS)-based unequal loss protection across source packets in packet erasure channels to maximize the average number of correctly decoded source bits. The second algorithm provides a product

code protection, i.e., the combination of RS code and an outer cyclic redundancy check (CRC) code/inner rate-compatible punctured convolutional (RCPC) code, for transmission over wireless channels including flat-fading.

Robustness to Various Channel Conditions

It is known that digitally encoded video data is susceptible to errors during wireless transmissions. In the standard video encoders, pixel values are transformed into frequency-domain coefficients by using discrete cosine transform (DCT) or wavelet and then quantized the coefficients. The quantized coefficients are finally compressed by variable length coding (VLC), i.e., Huffman coding. In VLC, even a single bit error can cause the loss of entire video data. In wireless data transmissions, error correction schemes are usually used for reliable communications. However, they generally have an all-or-nothing behavior for error correction. It means each received packet is either entirely correct or discarded according to the number of bit errors in each received packet. If a received packet contains bit errors, it causes significant degradation in received video quality.

Hanzo (1998) investigated the impacts of modulation and channel coding schemes on the received video quality in additive white Gaussian noise (AWGN) and Rayleigh fading channels. From the evaluation results, wireless video delivery suffers significant quality degradation at a low channel quality due to many bit errors. This phenomenon is called cliff effect. In addition, the video quality does not improve even when the wireless channel quality is improved. This is called leveling effect. This is because quantization is a lossy process, and its distortion cannot be recovered at the receiver. Cherriman et al. (1999) also investigated the impacts of frequency-selective fading on the received video quality. Here, they use Orthogonal Frequency-Division Multiplexing (OFDM) for dispersive channels to convey high-rate video contents.

Ji et al. (2003) solved the problems of video broadcasting over the OFDM-MIMO systems. In OFDM-MIMO systems, it takes advantage of both layered video coding and OFDM-

MIMO subchannels by matching the source layers of higher importance into the spatial subchannels with higher channel gain. The layered video source should be encoded into multiple descriptions (replicas) for OFDM-MIMO broadcast to guarantee consistent performance over various subchannel gains. Here, FEC code-based multiple description coding (MDFEC) is applied to the layered video source for multiple description coding. Ahmad et al. (2016) designed a cross-layer framework for multiuser wireless video streaming under the consideration of time-varying interference. A main challenge is to realize joint power control and rate adaptation to provide fair streaming quality across the users in time-varying wireless channels. They modeled the problem as a stochastic control problem and proposed a risk-sensitive control approach considering a nonlinear cost function to provide the optimal solution.

Scalability for Multiple Wireless Devices

When multiple users request the same video content/different video contents within the wireless links at the same time, it needs to share the same spectrum across the multiple wireless devices. In this case, a collision will occur when two or more wireless devices try to access the same spectrum at the same time. To efficiently share the spectrum, medium access control (MAC) protocols have been proposed for wireless communications. MAC protocol regulates channel access across multiple wireless devices to share the spectrum effectively and fairly. MAC protocols are usually classified into two types: contention-based and contention-free controls. The contention-based MAC protocol is a distributed approach that shares the spectrum among competitive devices. The contention-free MAC protocol brings the network up from a competitive state (where collisions may occur) to a collision-free state. The contention-free MAC protocols are often used in wireless video delivery because they can guarantee collision-free delivery.

On the other hand, with the growing demand for video applications, the MAC protocols need

to support various video applications with various Quality of Service (QoS) requirements. Specifically, real-time video applications require delay-sensitive controls, while video on demand (VoD) and multimedia surveillance applications need to support continuous connections between the sender and the receiver.

Other solutions to efficiently share the spectrum across multiple wireless devices are resource allocation of wireless medium and scheduling across the devices. Huand et al. (2008) considered efficient multiuser video streaming over wireless networks. A main objective is to maximize the received video quality of a limited number of video users, without interrupting the service of the coexisting voice users. For this purpose, they proposed a joint framework of network resource allocation, source adaptation, and deadline-aware scheduling for wireless multiuser video streaming. Specifically, they developed a joint algorithm, which consists of wireless resource allocation based on the video user's utility function, source rate adaptation based on content summarization, and packet scheduling to satisfy each video user's deadline constraints. Ji et al. (2009) solved the problems of scheduling and resource allocation for multiuser scalable video streaming over downlink OFDM systems. They proposed a scheduling and resource allocation algorithm across multiple users based on the users' priority weights calculated by the video contents, deadline requirements, and transmission history to thoroughly explore the time, frequency, and multiuser diversities of the OFDM system. He and Liu (2014) designed a cross-layer scheme with the packet scheduling and resource allocation algorithms to maximize the received video quality with the limited network resource and the deadline constraints over Orthogonal Frequency-Division Multiplexing Access (OFDMA) network. The proposed scheme considers the information extracted from the application layer, MAC layer, and physical layer and formulates the packet scheduling, subcarrier assignment, and power allocation into a mathematical model.

Recent Viewpoints in Wireless Video Delivery

Even since the 2010s, novel cross-layer techniques on wireless video delivery, including soft video delivery and video-application-aware MAC protocols, have been proposed under the considerations of video contents and wireless channel features to provide better video quality at wireless devices.

Soft Video Delivery

In conventional video delivery, they use VLC to compress the quantized frequency-domain coefficients and send the compressed bits to wireless devices. If bit flips occur by bit errors during wireless video delivery, it causes synchronized problems in video decoding and significant quality degradation. In addition, different wireless devices experience quite different wireless channel qualities, and each channel quality is usually unpredictable in practice. It means optimized quality adaptation according to each device's channel quality is not easy and sometimes suffer from cliff effect.

To realize better quality adaptation without actual channel information, Jakubczak and Katabi (2011) proposed soft video coding and decoding for wireless video delivery. In soft video delivery schemes, they use linear analog coding instead of VLC for compression. The analog coding mainly consists of three-dimensional DCT (3D-DCT), power allocation, and Walsh-Hadamard transform. 3D-DCT compacts video information by using spatial- and time-domain correlations between video frames. The power allocation scales each DCT coefficient according to the significance of the DCT coefficient to protect from effective noise. Walsh-Hadamard transform randomizes all the scaled coefficients. As a result, all packets have equal energy, i.e., significance, and thus realize better loss resilience.

By using analog coding for wireless video delivery, it can prevent significant quality degradation due to bit errors and packet losses, instead, realize graceful quality improvement with the improvement of wireless channel quality. Soft

video delivery schemes remove variable length digital coding for compression to prevent significant quality degradation due to synchronization problems.

Video-Application-Aware MAC Protocols

Video-aware MAC protocols have been proposed to satisfy various QoS requirements in wireless video applications. Specifically, the existing MAC protocols designed for delay-sensitive and quality-sensitive applications.

Wang and Zhuang (2009) and Chen et al. (2010) proposed contention-free and contention-based MAC protocols, respectively, under a certain delay constraint. In Wang and Zhuang (2009), they designed time-division multiple access (TDMA)-based MAC protocol for wireless video applications. Specifically, the time is partitioned into multiple slots of constant duration. Here, each slot consists of two-portions: control part and transmission part. The control part also consists of one or more real-time minislots and multiple non-real-time minislots. Each sender is allocated to each non-real-time minislot and thus sends a jamming signal at the corresponding minislot if the sender wants to send a packet. If the packet has a certain delay constraint, the sender also sends a jamming signal at the real-time minislot to guarantee QoS requirements of video applications.

In Chen et al. (2010), they proposed a content-aware retry limit adaptation scheme at the MAC layer for wireless video delivery. In IEEE 802.11-like networks, they use a retry mechanism, in which a lost packet is retransmitted several times up to a predefined retry limit. This may cause quality degradation because a packet may arrive later than the presentation time. The proposed scheme adaptively decides the retry limit of each packet based on the loss impact. Specifically, they calculate the amount of decoding errors for each packet when the packet is lost. If the amount of decoding errors is large, they assign high retry limit to decrease the impact of packet loss on the received video quality.

Baik et al. (2015) proposed contention-based MAC protocol considering the perceptual video quality into frame retransmission strategy. They

first figured out an impact of frame loss at MAC layer on visual quality. Based on the investigation, they adaptively decide the retransmission at MAC layer. For example, even when a transmitted frame is lost, a sender does not retransmit the frame when the user will accept the received video quality. By reducing the number of retransmissions under the consideration of perceptual video quality, it maintains the quality of wireless video applications with the reduction of network delays.

Key Applications

High-quality, reliable, and scalable wireless video delivery is required for low-resolution video contents as well as high-resolution and immersive video contents, e.g., multi-view video, free viewpoint video, and 360-degree video (Fujihashi et al. 2019). For example, multi-view video is an emerging and attractive technique to observe a three-dimensional (3D) scene from freely switchable angles. Multi-view video consists of multiple cameras' video captured by closely spaced camera arrays. A sender conveys video frames of multiple cameras over wireless links/networks, and each wireless user freely switches the viewing camera during video playback to figure out the preferred viewpoint of the scene.

Although main applications of wireless video delivery are downlink, uplink-based high-quality wireless video delivery is also required for recent wireless video applications such as crowdsourced video applications and video surveillance (Nu et al. 2018). In crowdsourced video applications, users can share and report the experience with other users by uploading video contents captured at a crowded event.

Cross-References

- [Scalable Video Streaming](#)
- [Wireless video Streaming](#)

References

- Ahmad A, Hassan NU, Assaad M, Tembine H (2016) Joint power control and rate adaptation for video streaming in wireless networks with time-varying interference. *IEEE Trans Veh Technol* 65(8):6315–6329
- Baik E, Pande A, Mohapatra P (2015) Efficient MAC for real-time video streaming over wireless LAN. *ACM Trans Multimed Comput Commun Appl* 11(4):1–24
- Chen CM, Lin CW, Chen YC (2010) Cross-layer packet retry limit adaptation for video transport over wireless LANs. *IEEE Trans Circuits Syst Video Technol* 20(11):1448–1461
- Cherrihan P, Keller T, Hanzo L (1999) Orthogonal frequency-division multiplex transmission of H.263 encoded video over highly frequency-selective wireless networks. *IEEE Trans Circuits Syst Video Technol* 9(5):701–712
- Fujihashi T, Akino TK, Watanabe T, Orlik P (2019) FreeCast: graceful free-viewpoint video delivery. *IEEE Trans Multimedia* 99:1–11
- Hanzo L (1998) Bandwidth-efficient wireless multimedia communications. *Proc IEEE* 86(7):1342–1382
- He L, Liu G (2014) Quality-driven cross-layer design for H.264/AVC video transmission over OFDMA system. *IEEE Trans Wirel Commun* 13(12):6768–6782
- Hsu CY, Ortega A, Khansari M (1999) Rate control for robust video transmission over burst-error wireless channels. *IEEE J Sel Areas Commun* 17:756–773
- Huang J, Li Z, Chiang M, Katsaggelos A (2008) Joint source adaptation and resource allocation for multi-user wireless video streaming. *IEEE Trans Circuits Syst Video Technol* 18(5):582–595
- Jakubczak S, Katabi D (2011) A cross-layer design for scalable mobile video. In: *ACM annual international conference on mobile computing and networking*, pp 289–300
- Ji Z, Zhang Q, Zhu W, Lu J, Zhang YQ (2003) Video broadcasting over MIMO-OFDM systems. In: *International symposium on circuits and systems*, pp 844–847
- Ji X, Huang J, Chiang M, Lafruit G, Catthoor F (2009) Scheduling and resource allocation for SVC streaming over OFDM downlink systems. *IEEE Trans Circuits Syst Video Technol* 19(10):1549–1555
- Khansari M, Jalali A, Dubois E, Mermelstein P (1994a) Robust low bit-rate video transmission over wireless access systems. In: *Smith J (ed) International conference on communications (ICC/SUPERCOMM)*, New York, pp 571–575
- Khansari M, Zakaudinn A, Chan WY, Dubois E, Mermelstein P (1994b) Approaches to layered coding for dual-rate wireless video transmission. In: *International conference on image processing*, pp 1–5
- Nu TT, Fujihashi T, Watanabe T (2018) A traffic reduction method for crowdsourced multi-view video uploading. *IEEE Access* 6:36544–36556
- Ortega A, Khansari M (1995) Rate control for video coding over variable bit rate channels with applications to wireless transmission. *IEEE Int Conf Image Process* 3:388–391
- Pelz RM (1994) An unequal error protected px8 kbits/s video transmission for DECT. In: *IEEE vehicular technology conference*, pp 1020–1024
- Schwarz H, Marpe D, Wiegand T (2007) Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Trans Circuits Syst Video Technol* 17(9):1103–1120
- Sheng S, Chandrakasan A, Brodersen RW (1992) A portable multimedia terminal. *IEEE Commun Mag* 30(12):64–75
- Song D, Chen CW (2007) Scalable h.264/avc video transmission over MIMO wireless systems with adaptive channel selection based on partial channel information. *IEEE Trans Circuits Syst Video Technol* 17(9):1218–1226
- Stankovic V, Hamzaoui R, Xiong Z (2004) Real-time error protection of embedded codes for packet erasure and fading channels. *IEEE Trans Circuits Syst Video Technol* 14(8):1064–1072
- Wang P, Zhuang W (2009) A collision-free MAC scheme for multimedia wireless mesh backbone. *IEEE Trans Wirel Commun* 8(7):3577–3589

Wireless Video Streaming

Jie Li, Ransheng Feng, and Qiyue Li
School of Computer and Information, Hefei
University of Technology, Hefei, Anhui, China

Synonyms

[Video streaming in wireless networks](#)

Definitions

Wireless video streaming is a kind of real time video delivery pattern over wireless networks.

Historical Background

Video streaming is one of the most popular applications for wireless and mobile networks. Starting from the mid-1990s, many works focused on performance analysis for real time video delivery over wireless networks. A large amount of attentions was given to enhance video coding in order to combat errors introduced by the wireless chan-

nel variability. Stuhlmüller et al. (2000) derived a theoretical framework for the picture quality after video transmission over lossy channels, based on a 2-state Markov model describing burst errors on the symbol level. Zhang et al. (2000), Li et al. (2016), Liu et al. (2018), and He et al. (2002) proposed methods for estimating the channel distortion and its impact on performance.

Video streaming over wireless networks is increasingly the technology of choice for a wide range of important applications, including remote surveillance, security monitoring, smart home, environmental tracking, and battlefield intelligence. A major challenge in wireless streaming is that the channel is time-varying and error-prone. In video streaming over wireless channels, each video packet can be received correctly, received with errors, or lost over the network. If a video packet is received with errors, the error bits will cause decoding failure of the packet. In this case, the video decoder often discards the packet and jumps to the next one to continue the decoding process. Major causes of packet loss include network congestion and buffer overflow. In real-time applications, video transmission has a delay constraint, i.e., the video packet has to arrive at the decoder before the scheduled playback time. In video transmission over shared wireless channels, packet delivery often experiences significant delays. These works mainly focused on ensuring robustness of video delivery over a variable wireless channel.

Streaming media over wireless networks is becoming an increasingly important application. An Unequal Error Protection (UEP)-based error control scheme for transmission of low bit rate MPEG-4 video over wireless channels is proposed, where MPEG-4 bitstream is divided into two classes, and UEP is applied to each class of data.

Later on, there has been a tremendous growth in streaming media applications, thanks to the fast development of network services and the remarkable growth of smart mobile devices. HTTP Live Streaming (HLS) (Apple Inc), Silverlight Smooth Streaming (MSS) (Zambelli), HTTP Dynamic Streaming (HDS) (Adobe Systems Inc), and Dynamic Adaptive

Streaming over HTTP (DASH) (DASH Industry Forum) achieve decoder-driven rate adaptation by providing video streams in a variety of bitrates and breaking them into small HTTP file segments. The media information of each segment is stored in a manifest file, which is created at server and transmitted to client to provide the specification and location of each segment. Throughout the streaming process, the video player at the client adaptively switches among the available streams by selecting segments based on playback rate, buffer condition, and instantaneous TCP throughput (Stockhammer 2011). Specially, Zhang et al. (2016) and Liu et al. (2014) study the influence of network coding of free viewpoint video for wireless streaming.

The video sources for most streaming applications are typically precoded video sequences with relative high bit rates. However, the currently deployed wireless cellular systems are designed to only support lower bit rate data. In order to support video streaming over such networks, the high rate video sources need to be processed through a variety of schemes, such as scalable video stream extraction, transcoding, and summarization before they can be accommodated by the wireless channel. Different video content segments have different rate-distortion characteristics, e.g., some segment may be part of an action movie and requires a lot of bits to encode, while others may be just static video frames that require relatively less bits. In a wireless multiaccess channel, the type of multiuser content diversity in content rate-distortion characteristics should be taken into consideration while optimizing the network resource.

In summary, wireless video streaming has great application value in the field of streaming media.

Foundations

For Wireless video streaming, in general, media players at the devices are equipped with a playout buffer that stores arriving packets. As long as there are packets in the buffer, the video can

be played smoothly. However, starvations of the buffer can cause large jitters and are particularly annoying for end users that see frozen images. One feasible way to avoid starvations is to introduce a start-up (also called prefetching) delay before playing the stream and a rebuffering delay after each starvation event. Then after a number of media frames accumulate in the buffer, the media player starts to work.

Indeed, wireless channel is subject to a large variability due to fading, mobility, etc. For example, in 3GPP LTE cellular network, receiver bandwidth is defined as a certain number of Physical Resource Blocks (PRB) in both time and frequency domain. The more PRBs are assigned to one user, the more bitrate he can enjoy. In Sallabi and Shuaib (2011), a Finite State Markov Channel (FSMC) model is proposed to estimate the bandwidth of LTE wireless network. The states are determined by partitioning the average received Signal to Noise Ratio (SNR) range to K intervals, where $K-1$ is the number of Modulation Coding Scheme (MCS). The variation of PRBs can be modeled as a stationary Markov chain, with state space $S = \{S_0, S_1, \dots, S_M\}$. The state space contains K different PRB states with corresponding bit per symbol rate; each state records the number of RRB available. Let π_i is the probability of stationary state i and π_i is the onestep transition probability from state i to state j , where $i, j \in \{0, 1, \dots, K-1\}$. The model is illustrated in Fig. 1. Here we give an example to further explain this; the MCSs recommended by 3GPP are QPSK, 16QAM, and 64QAM. These three MCSs lead to four intervals and four states. We can estimate the real-time bandwidth of wire-

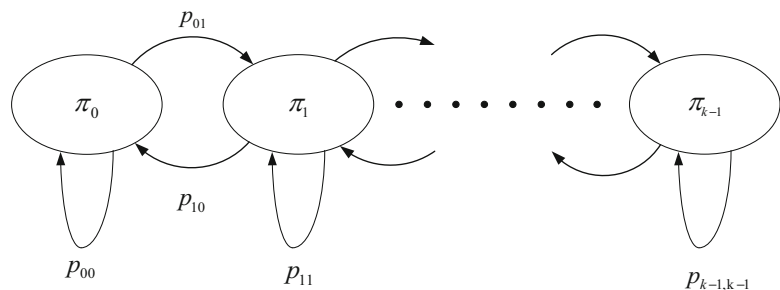
less channel based on this model. Then whether the video streaming will stall or not can be known by comparing the available bandwidth with the video bitrate.

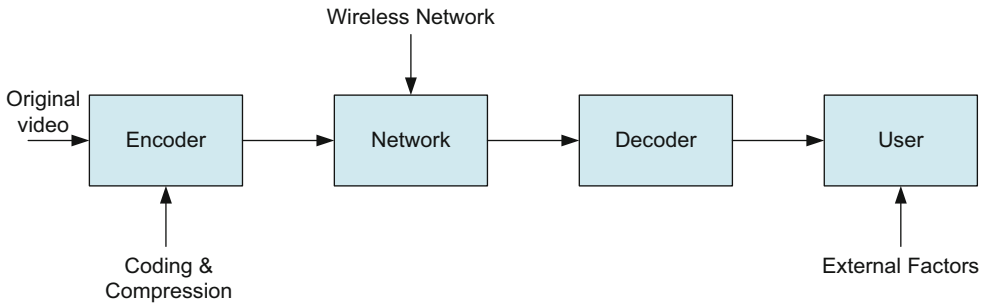
Streaming video services can be classified into three major types: Real-Time Streaming Service (RTSV), Near-Real Time Streaming Service (NRTSV), and Download service (R1-083233 2008). The RTSV service type is suitable for interactive video conferencing applications. The primary constraint for this type of service is guarantees on the end-to-end latency of the video frame transmission.

The NRTSV service is suitable for noninteractive streaming applications, which can be either live streaming or streaming of archived media. At the beginning, the video frames of the stream are buffered into a preroll buffer, or de-jitter buffer. The playout begins after the buffering delay, which is typically several seconds, e.g., 5 s to 30 s. The primary guarantee required for this service is that there is always a video frame available for playout in the preroll buffer after the playout begins.

The Download service is suitable when the mobile terminal has a large storage available to download and save the video stream, which can be played out later. There are two types of Download services possible: complete download service (also known as cache and carry) and progressive download service. In complete download, the entire video stream is downloaded and stored into the mobile device; it can be played at a later time at the user's discretion. In the progressive download service, the mobile device starts playout after downloading a fraction of the

Wireless Video Streaming, Fig. 1 FSMC model with k states





Wireless Video Streaming, Fig. 2 Wireless video streaming flow

entire video content file; then the playout and download continues concurrently.

The wireless video streaming flow from the server side to the client side includes: encoder, network, decoder, and display, as shown in Fig. 2.

- *Coding and compression:* In order to transmit rich video content through a bandwidth limited network, the original video information needs to be compressed. Video compression formats define the way to represent the video and audio as a file or a stream. Video codec, a device or software, encodes or decodes a digital video based on the video compression format.
- *Network:* Common transmission network that is considered in the video streaming is wireless network. For the wireless network, it includes cellular network, wireless local area network (WLAN), vehicular network, and sensor network. Transmission network condition will greatly affect the video quality. There are three major factors that will influence video quality: TCP throughput, which represents the bandwidth; delay, which depends on the network capacity; and packet loss, which is due to unreliable transmission.
- *External factors:* Apart from the above factors, there are some external factors influencing wireless video streaming. The following are external factors: video service type, whether the video is live streaming video or Video-on-Demand; video length, It has been proven that users have different viewing

behaviors for videos of different lengths; and video popularity, generally, people prefer to watch popular videos.

Recently, a new class of video transport techniques has been introduced for transmission of video over varying channels such as wireless network. These transport techniques, called adaptive streaming, vary the bit rate and quality of the transmitted video to match the available channel bandwidth and alleviate the problems caused by network congestion, such as large latency and high packet loss rate. DASH, Dynamic Adaptive Streaming over HTTP, is a new international standard for adaptive streaming (Dynamic Adaptive Streaming Over HTTP 2010), which enables delivering media content from conventional HTTP web servers. DASH works by splitting the media content into a sequence of small segments, encoding each segment into several versions with different bit rates and qualities, and streaming the segments according to the requests from streaming client. On the client device side, the DASH client will keep monitoring the network and dynamically select the suitable version for the next segment that need to be downloaded, depending on the current network conditions.

Key Applications

Wireless video streaming is an efficient and fundamental video transmission pattern in streaming

media. There are many applications of wireless video streaming in the field of streaming media. In terms of real-time video streaming, most are related to wireless video streaming. For example, live video and video conferencing are all applications of wireless video streaming. Common transmission networks that are considered in the video streaming include IP network and wireless network. And for the wireless video streaming, the network it uses is a wireless network which includes cellular network, wireless local area network (WLAN), vehicular network, and sensor network.

Future Directions

For wireless video streaming, there are three aspects that need to be improved in the future:

- *Improve transmission channel quality.* Transmission network condition will greatly affect the video quality in the wireless video streaming. There are three major factors that will influence video quality: TCP throughput, delay, and packet loss. TCP throughput influences the bandwidth of transmission, the later cause frame loss in wireless video streaming. Thus, we should improve the quality of transmission channel in the future.
- *Reduce interference on the transmission channel.* The channel will change continuously due to various influencing factors during transmission. This phenomenon will greatly affect the quality of the video in the wireless video streaming. Thus, we must minimize the interference of the channel's time-varying video quality.
- *Improve video coding efficiency.* Improve video coding efficiency can increase the transmission quality of video to a certain extent. So we should improve it in the future.

Cross-References

- [Wireless Network](#)

References

- Adobe Systems Inc. HTTP dynamic streaming 2013, online. <http://www.adobe.com/products/hds-dynamic-streaming.html>. Accessed 15 Feb 2015
- Apple Inc. HTTP live streaming technical overview 2013, online. <http://developer.apple.com/library/ios/documentation/networkinginternet/conceptual/streamingmediaguide/Introduction/Introduction.html>. Accessed 15 Feb 2015
- DASH Industry Forum. For promotion of MPEG-DASH 2013, online. <http://dashif.org>. Accessed 12 Feb 2015
- Dynamic Adaptive Streaming Over HTTP, w11578, ISO/IEC JTC 1/SC 29/WG 11 Standard CD 23001-6, 2010
- He ZH, Cai JF, Chen CW (2002) Joint source channel ratedistortion analysis for adaptive mode selection and rate control in wireless video coding. *IEEE J Sel Areas Commun* 12(6): 511–523
- Li J, Bao Z, Zhang C, Li Q, Liu Z (2016) Joint mcs and power allocation for svc video multicast over heterogeneous cellular networks. *Comput Commun* 83(C):16–26
- Liu Z, Cheung G, Chakareski J, Ji Y (2015) Multiple description coding and recovery of free viewpoint video for wireless multi-path Streaming. *IEEE J Sel Top Signal Processing*. 9(1):151–164
- Liu Z, Ishihara S, Cui Y, Ji Y, Tanaka Y (2018) Jet: joint source and channel coding for error resilient virtual reality video wireless transmission. *Signal Process* 147:154–162
- RI-083233, “Video Services over LTE-Advanced”, Motorola, 3GPP TSG RAN1#54, Jeju, Korea, 18–22 Aug 2008
- Sallabi F, Shuaib K (2011) Modeling of downlink wireless fading channel for 3gpp lte cellular system. In: 2011 IEEE GCC conference and exhibition (GCC), pp 421–424
- Stockhammer T (2011) Dynamic adaptive streaming over http-: standards and design principles. *Proc ACM Conf Multimedia Syst*: 133–144
- Stuhlmüller K, Farber N, Link M, Girod B (2000) Analysis of video transmission over lossy channels. *IEEE J Sel Areas Commun* 18(6):1012–1032
- Zambelli A (2009) Smooth streaming technical overview, online. <http://www.iis.net/learn/media/on-demand-smooth-streamingsmoothstreamingtechnical-overview>. Accessed 15 Feb 2015
- Zhang R, Regunathan SL, Rose K (2000) Video coding with optimal inter/intra mode switching for packet loss resilience. *IEEE J Sel Areas Commun* 18(6): 966–976
- Zhang B, Liu Z, Gary Chan SH, Cheung G (2016) Collaborative wireless freeview video streaming with network coding. *IEEE Tran Multimedia*. 18(3): 521–536

Wireless Video Transmission

- ▶ [Wireless Video Delivery](#)

WSN Frameworks

- ▶ [Middleware for Wireless Sensor Networks](#)

WLAN Authentication

- ▶ [Authentication](#)

WSN Management

- ▶ [Middleware for Wireless Sensor Networks](#)

Z

Zero-Rating

- [Game Theoretic Modeling of Zero-Rating Internet Provisioning](#)